Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ 6/24/20_____

Cristina E. Trevino Date

Maximizing the ability to detect modifying genetic factors of rare complex disorders – Fragile X-Associated Primary Ovarian Insufficiency and Down Syndrome - Congenital Heart Defects

By
Cristina E. Trevino
Doctor of Philosophy

Graduate Division of Biological and Biomedical Science Genetics and Molecular Biology

_____
Stephanie L. Sherman,
Ph.D. Advisor

_____
David J. Cutler,
Ph.D. Committee Member

_____
Michael P. Epstein,
Ph.D. Committee Member

_____
Judith L. Fridovich-Keil,
Ph.D. Committee Member

_____
Peng Jin,
Ph.D. Committee Member

Accepted:

_____

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____ Date

Maximizing the ability to detect modifying genetic factors of rare complex disorders – Fragile X-Associated Primary Ovarian Insufficiency and Down Syndrome - Congenital Heart Defects

By

Cristina E. Trevino

B.S, Georgia Institute of Technology, 2012

Advisor: Stephanie L. Sherman

An abstract of
A dissertation submitted to the Faculty of the James T. Laney School of Graduate Studies of Emory University in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in
Graduate Division of Biological and Biomedical Science
Genetics & Molecular Biology
2020

Maximizing the ability to detect modifying genetic factors of rare complex disorders – Fragile X-Associated Primary Ovarian Insufficiency and Down Syndrome - Congenital Heart Defects

Abstract

In order to better identify and understand the genetic architecture of complex traits, modern genomic methods are more focused on using the ample amount of data that has been collected over the last decade and examining the genome in different ways. However, prioritizing functional variants in this framework remains challenging. Strategies including faster and easier to use annotation and filtering methods are increasingly important for genomic analyses today. Selecting cohorts from genetically-sensitized populations or constructing a cohort from those with the extreme phenotypes of a complex trait are other strategies to maximize the ability to detect susceptibility variants. In this dissertation, I employ these strategies to study primary ovarian insufficiency (POI) in a cohort of women with a fragile X premutation (PM) and to study atrioventricular septal defects (AVSD) in a cohort of individuals with Down syndrome (DS). Both of these groups have these co-occurring traits at a much higher frequency than the general population - women with a PM are at a 20-fold increased risk for POI and individuals with DS are at a >2000 increased risk for AVSD.

POI, which affects 1% of women in the general population, is a condition characterized by symptoms of early menopause and is a leading cause of infertility. About 20% of women who carry a PM, a CGG repeat expansion in the range of 55-200 repeats in the 5'UTR of the X-linked *FMR1* gene, are diagnosed with fragile X-associated POI (FXPOI). We hypothesize that there are genetic modifiers that contribute to the age of onset and severity of FXPOI. In order to test this, we conducted a case/control study among women with a PM taken from the extremes of the distribution of age at onset of FXPOI/menopause (onset before age 35 and after age 50). We compared whole genome sequencing (WGS) data in an untargeted way and examining candidate genes that are involved in the underlying mechanism of PM-associated disorders.

Top ranked genes were then screened using the *Drosophila* model as a high-throughput, whole organism functional screen to gain further evidence of their involvement in ovarian dysfunction.

AVSDs are a rare and severe form of congenital heart defects (CHD) and require surgery soon after birth. In general, CHDs occur in almost 1% of infants in the general population; AVSD occurs in about 1/10,000. Most genetic studies of CHD examine all forms, although there is strong evidence of etiological heterogeneity. We took the same strategy as above and identified a genetically-sensitized population to increase the ability to identify risk variants of AVSD. About 20% of infants with Down syndrome, or trisomy 21, are born with an AVSD, an enormous increase in frequency over the general population. Thus, we based our study on 702 individuals with DS who did and did not have an AVSD, again, drawing from those with the extremes of heart development. We used available whole exome sequencing, WGS, and/or array-based imputation data and took a variety of statistical approaches to examine risk-associated genes and pathways and to examine the contribution of many common variants of small effect size using polygenic risk score (PRS) methods.

Results from both studies that combined multiple statistical approaches of genetic data based on extreme phenotypes within genetically-sensitized cohorts proved successful. Identified candidate genes can now be moved to mammalian model systems to test for functional involvement. These studies benefit not only those with increased risk (i.e., women with a PM or people with DS), but may also be translated to those with idiopathic forms of the disorders.

Maximizing the ability to detect modifying genetic factors of rare complex disorders – Fragile X-Associated Primary Ovarian Insufficiency and Down Syndrome - Congenital Heart Defects

By

Cristina E. Trevino

B.S, Georgia Institute of Technology, 2012

Advisor: Stephanie Sherman

A dissertation submitted to the Faculty of the James T. Laney School of Graduate Studies of Emory University in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in
Graduate Division of Biological and Biomedical Science
Genetics & Molecular Biology
2020

Abstract
Table of contents

III.III. Methods
- III.III.i. Subjects
- III.III.ii. Whole exome sequencing
- III.III.iii. Whole genome sequencing
- III.III.iv. Samples with imputed genotypes based on microarray
- III.III.v. SKAT-O variant analysis
- III.III.vi. Polygenic risk score analyses
  - III.III.vi.i.Target dataset for primary analyses
  - III.III.vi.ii.Target dataset for secondary PRS analyses
  - III.III.vi.iii.Generating PRS for the primary analyses
  - III.III.vi.iv.Generating PRS for the secondary analyses
  - III.III.vi.iv.Testing association of PRS with DS+AVSD

III.IV. Results
- III.IV.i. Gene discovery using SKAT analyses
- III.IV.ii. CHD polygenic risk score and its association with DS+AVSD
  - III.IV.ii.i Primary analyses indicate a non-significant association of the CHD-based PRS with DS+AVSD
  - III.IV.ii.ii Adding data from chromosome 21 into the PRS calculation did not change the association with DS+AVSD

III.V. Discussion
III.VI. Tables and Figures
II.VII. References
II.VIII. Supplemental methods and References

IV. Bystro: rapid online variant annotation and natural-language filtering at whole-genome scale
- IV.I. Abstract
- IV.II. Introduction
- IV.III. Results
- IV.IV. Discussion
- IV.V. Methods
  - III.V.i. Accessing Bystro
  - III.V.ii. Bystro Database
  - III.V.iii. WGS Datasets
  - III.V.iv. Online annotation comparisons
  - III.V.v. Variant filtering comparisons
  - III.V.vi. Filtering accuracy comparison
  - III.V.vii. Offline annotation comparisons
  - III.V.viii. Annotation accuracy comparison
- III.VI. Tables and Figures
- II.VII. References

V. Discussion
- V.I Conclusions
- V.II Limitations
- V.III Implications and future directions

V.IV References

Tables

Figures

I. **Introduction**

I.I Understanding the genetic architecture of complex traits

Identifying genetic factors that contribute to complex traits in the past decade has been primarily accomplished through genome-wide association studies (GWAS) (Visscher et al., 2017). All of the data gathered from these studies have been catalogued and are now available to use to further the research in different ways (Buniello et al., 2019). Similarly, data from large whole exome sequencing (WES) and whole genome sequencing (WGS) studies are now deposited in public, access-controlled databases (e.g., dbGaP) for general use. Projects like the ENCODE project have continuously helped to bring annotated information from these genomic studies together to make it easier to access (Luo et al., 2020). Recent developments in statistical methods in genomics have been expanding more on the idea of grouping both common and rare variants into genes and pathways as an analytical unit rather than individual variants to better characterize the profile of genetic diseases and to obtain more power in genomic analyses (Moutsianas et al., 2015). Better genomic annotation methods facilitate the speed of these analyses and improved statistical methods to identify susceptibility variants bolster future studies overall. In addition to the improvement of genomic tools, annotation and statistical tools, new study designs are being considered to increase the ability to understand the genetic architecture of rare disorders.

In this dissertation, I have taken advantage of cohorts of individuals from more sensitized populations to detect susceptibility variants that can potentially be translated

to those affected in the general population. More specifically, this strategy can be implemented in cohorts of individuals with rare disorders with a large-effect background mutation where co-occurring conditions are seen at a much higher rate than the general population. Examples include congenital heart defects in Down Syndrome and POI in women with the fragile X premutation (FXPOI). As we continue to work to identify the genetic background underlying complex disorders, these strategies will provide new information built on the resources generated from GWAS and gene annotation information.

I.I.i. Gene set analyses

To increase power for rare variant analyses, variants may be filtered to consider only those exceeding a certain severity threshold of predicted deleteriousness, and then grouped by gene or pathway and tested simultaneously in these variant sets. Burden tests are often used for this purpose, however a burden test assumes variants are contributing to a phenotype in the same direction - either protective or risk-associated.

For biological phenotypes, this assumption may not be warranted as variants could be acting in either direction. In order to overcome this limitation, the sequence kernel association test (SKAT) may be applied (Wu et al. 2010), which allows for modeling of the joint effect of risk and protective alleles within a set via a logistic kernel-machine-based test that can also include covariates. An optimal unified test (SKAT-O) maximizes the value in both types of combined variant testing (S. Lee et al. 2012) by modeling both SKAT and the burden test for each defined variant set and finding the optimal linear combination of both tests. Thus, it optimizes power for all scenarios.

SKAT-O also minimizes type I errors in smaller sample sizes by estimating the sample variance and kurtosis allowing for proper reference distribution (S. Lee et al. 2012). As well as rare variant testing, SKAT-O can also be employed for analysis of common variants within genes and pathways.

I.I.ii. Understanding contribution of polygenes

Extensive work has been done examining and cataloguing the common variants in complex traits (Buniello et al., 2019). One of the powerful tools that is now available given this body of work is being able to examine polygenic risk scores (PRS) in human genetics using public GWAS data. PRS are calculated as a weighted sum of risk alleles in which the weights are generated from a GWAS of the complex trait being measured as a discovery sample (Chatterjee et al., 2016). Thus far PRS analyses have primarily been used to better understand polygenic risk that was previously challenging to determine in various complex disorders (Chalmer et al., 2018; Cleynen et al., 2020; Escott-Price et al., 2017; Kauppi et al., 2015). PRS profiling is also being considered to determine if genome-wide polygenic risk can be used to improve screening for and identify individuals with genetic risk for common disorders (Khera et al., 2018; Torkamani et al., 2018). This approach to better understand underlying heritability and genetic risk is limited by the available GWAS data. In addition to this limitation, even when there are available GWAS data, the amount of variability that can be explained for any given polygenic trait is limited by the cohort size of the discovery GWAS set, which will often be small for rare diseases.

I.II Fragile X Primary Ovarian Insufficiency

I.II.i. Prevalence of Primary Ovarian Insufficiency

Primary Ovarian Insufficiency (POI) is a leading cause of female infertility and is characterized by cessation of menses for at least four months before the age of 40 and increased follicle-stimulating hormone (FSH) levels > 25 IU/l measured twice (Rudnicka et al., 2018; Webber et al., 2016). POI is present at 1% in the general population (Nelson, 2009). Women with POI are also at a greater risk of disorders associated with early estrogen-deficiency, such as osteoporosis, type 2 diabetes, cardiac disease, and all-cause mortality (Anagnostis et al., 2019; Jacobsen et al., 2003; Muka et al., 2016; Shuster et al., 2010). Multiple genes and chromosomal abnormalities including but not limited to Turner Syndrome/monosomy X, *GALT*, *CHM, DIAPH2, POF1B, XPNPEP2, NXF5, USP9X, ZFX, BMP15, FMR1, FMR2, XIST, CENPI, PGMRC1, AR, FOXO4, AGTR2 and BHLHB9* have been associated with risk of POI (Fortuño & Labarta, 2014; Goswami & Conway, 2005). Despite these associations, etiology is unknown in the majority of POI cases (Vujovic, 2009). One important cause of POI is the fragile X premutation (PM) allele, a CGG repeat expansion in the 5'UTR of the FMR1 gene (Figure 1.1). Indeed, it is the most common single gene cause of POI. Among those with POI, the PM is identified in about 11% of women with a family history of POI and about 3% among isolated cases (Bussani et al., 2004; Marozzi et al., 2000; Murray, 2000). The frequency of women who carry a PM allele in the general population is approximately 1/300 (Cronister et al., 2005; Hagerman, 2008; Hantash et al., 2011;

Lévesque et al., 2009; Song et al., 2003). Among those women, the risk of POI is about 20%, thus a 20-fold increased risk compared with the general population (Nelson, 2009; S. L. Sherman, 2000).

I.II.ii. Risk factors for FXPOI

Although there is a significant increase of POI among women with a PM, the majority of carriers go through menopause around 50 years of age. Several genetic and environmental factors have been studied to determine if they contribute to the incomplete penetrance in FXPOI. The repeat size of a PM itself confers risk non-linearly, with the highest risk occurring at 80-100 repeats (Allen et al. 2007; Sullivan et al. 2005; Spath et al. 2011; Ennis, Ward, and Murray 2006). An additive genetic component separate from the PM that confers risk to FXPOI has been evaluated (Hunter et al., 2008; Spath et al., 2011). However, the specific modifying genes underlying the genetic background contribution to FXPOI outside of the PM have yet to be fully determined beyond studies that show there is a genetic background component contributing to incidence of FXPOI in women with the PM (Hunter et al., 2008).

Another genetic factor that has been evaluated for its contribution to incidence of FXPOI is the genetic component of age at natural menopause (AOM). Thus far, there have been several GWAS studies of women who experience natural AOM in which over 50 loci have been found to have an association with AOM (Day et al., 2015; Perry et al., 2014; Stolk et al., 2012). Many of the associated variants were found in DNA damage response genes (Day et al., 2015; Perry et al., 2014; Stolk et al., 2012). The primary

environmental factor considered for association with FXPOI is smoking, but any association has been found to be due to smoking impacting age at menopause for all women by reducing age at menopause (Allen et al., 2007; Spath et al., 2011). Genetic factors that have been determined to not be associated with risk of FXPOI are age at menarche, BMI, and skewed X-inactivation (Bione et al., 2006; Hunter et al., 2008; Rodriguez-Revenga et al., 2009; Spath et al., 2010; Tejada et al., 2008).

I.II.iii. Mechanisms of the PM leading to FXPOI

Mechanisms of Fragile X-associated disorders have primarily been discovered in studies of Fragile X-associated tremor/ataxia syndrome (FXTAS) (Aumiller et al., 2012; Sellier et al., 2010, 2013; Sofola et al., 2007; Todd et al., 2013). There are two primary mechanisms that have been identified for how the PM can lead to fragile X-associated disorders - repeat associated non-AUG (RAN) translation and sequestration of RNA binding proteins (Figure 1.2). RAN translation is a noncanonical process of translation in which translation machinery is stalled at a structure like the hairpin formed by the CGG repeats in the fragile X premutation and then translate small polypeptides, in this case polyalanine and polygycine products (Reddy & Pearson, 2013). RAN translation occurs in multiple repeat disorders (Ash et al., 2013; Mori et al., 2013; Zu et al., 2011). Sequestration of RNA binding proteins also involves the repeat mRNA forming structures but instead of generating a new genetic product, RNA binding proteins that directly bind CGGs and proteins bound to them are bound to the structures formed by

the repetitive mRNA, resulting in a loss of function for those proteins (Sellier et al., 2010, 2013).

Drosophila has been a useful model for studies of Fragile X tremor-ataxia syndrome (FXTAS) (Jin et al., 2003). In favor of the sequestration model where RNA binding proteins are sequestered by the PM repeat, this model showed that specific CGG RNA binding proteins alter neuronal function, including *hnRNP A2/A1*, *CUGPB1*, and *Pur-alpha* (Aumiller et al., 2012; Sofola et al., 2007). *Drosophila* models of other repeat disorders have also been used to study potential repeat toxicity from RAN translation products (Koon & Chan, 2017). Drosophila studies related to ovarian function have focused primarily on germline stem cells with respect to *FMR1*, in which they found *FMR1* played an important role in maintenance of germline stem cells and in repressing differentiation (L. Yang et al., 2007; Y. Yang et al., 2009).

I.II.iv. Animal models for FXPOI

Multiple animal models have been established for FXPOI (Reviewed in Sherman et al., 2014). Thus far, there have been multiple murine models in which different ovarian pathologies have been observed (Buijsen et al., 2016; Ferder et al., 2013; Hoffman et al., 2012; Lu et al., 2012). Using these PM models has been useful to understand ovarian function in the presence of the PM alone. For reasons of sample access, determining whether ovarian function is perturbed at different stages of development is only possible using animal models (Hoffman et al., 2012; Lu et al., 2012). In the study by Lu et al., 2012, ovarian size and follicle count were evaluated at postnatal days (PD)

8 and 25, and at 9 and 16 weeks. It was determined that there is no perturbation of early primordial follicle pool or ovarian size at PD25 compared to wildtype, but was a reduction in ovarian size and mature follicles at 9 weeks (Lu et al., 2012). A similar result was found in the study by Hoffman et al., 2012 at four months. Morphological discoveries that have come from studies of these models include depletion of follicles in later stages of life, granulosa cell abnormalities and increased atresia. Gene expression is also being studied in these mouse models, with alterations in the Akt/mTOR pathway (Buijsen et al., 2016; Lu et al., 2012). The findings in these models have already given a better understanding of the disease pathology that may underly FXPOI and continue to be essential in examining mechanisms for modifying genes that are found through genomic studies in women with the PM.

I.III Congenital Heart Defects in Down Syndrome

I.III.i. Prevalence and variability in phenotype for DS

Trisomy 21, also known as Down Syndrome (DS), is genetically complex and results in a variety of phenotypes including intellectual disability, congenital heart defects, and developmental delay (Antonarakis et al., 2020). Variation in phenotype can be attributed to many of the genes on the long arm of chromosome 21 (Antonarakis, 2017). Structural variants in this region as well as SNVs contribute to the global dysregulation of the transcriptome (Antonarakis, 2017). However, the trisomy and variants on chromosome 21 do not fully explain all variation seen in DS. Specifically, genetic variation on non-

chr21 loci has also been shown to be an important component in explaining variability in DS clinical phenotypes (Antonarakis, 2017; Brown et al., 2019; Hertzberg et al., 2010; Sailani et al., 2013).

Certain phenotypes are more homogenous in the DS population than you would find in other groups because of the strong background of a genetic risk (Antonarakis, 2017; Karmiloff-Smith et al., 2016). One of the strategies that can be implemented to better understand the architecture of complex genetic traits such as CHD is to look at a population like DS in which there is a greatly increased occurrence of that trait. Congenital heart defects (CHD) occur in 80 out of 10,000 live births and it is the most common birth defect (Reller et al., 2008). In individuals with DS, the rate of congenital heart defects is over 40 fold higher than in the general population (Hartman et al., 2011). Atrioventricular septal defects (AVSD), a severe CHD for which surgery around the time of birth is needed, is found in DS at a greater than 2,000 fold prevalence in DS compared to the general population (Hoffman et al., 2012; Mai et al., 2015; Sarısoy et al., 2018). AVSDs cover a wide spectrum of congenital heart defects that are characterized by structural defects allowing blood flow between the left and right atria and/or ventricles instead of the complete separation of left and right atria and/or ventricles needed for efficient delivery of oxygenated blood to the body (Craig, 2006).

I.III.ii. Genetic studies of CHD

Many subtypes of CHDs represent a spectrum of phenotypes, but cohorts are sometimes studied together because of the rareness of some of the specific defects

(Cordell, Bentham, et al., 2013; Pierpont et al., 2007). The largest genomic cohort of CHDs to date includes 9,727 cases and includes seven CHD subtypes (Hoang et al., 2018). Both SNVs and CNVs have been studied for the association with CHDs (Cordell, Bentham, et al., 2013; Soemedi et al., 2012). Studying trios with subtypes of CHDs together have revealed a genetic background of *de novo* mutations contributing about 10% of severe CHDs (Zaidi et al., 2013). Common genetic variation was implicated in a GWAS of Tetralogy of Fallot (TOF), the most common severe CHD subtype (Cordell, Töpf, et al., 2013). CNVs have also been associated with TOF including CNVs in loci overlapping two genes, *NOTCH1 and JAG1* (Greenway et al., 2009). Rare variant (MAF < 0.01) studies have revealed enrichment for missense variants in *NR2F2*, a gene that encodes a nuclear receptor that is part of a steroid hormone superfamily and plays a role in heart development (Al Turki et al., 2014; Lin et al., 2012; Pereira et al., 1999). Missense mutations in *CRELD1* have also been implicated in the pathogenesis of AVSD (Robinson et al., 2003).

I.III.iii. Genetic studies of DS CHD

Both common and rare variants have been examined for genetic contribution to CHD in DS. So far, these studies have not found any large-effect common SNVs or CNVs that surpass genome-wide significance even with adequately powered studies (Ramachandran, Zeng, et al. 2015; Ramachandran, Mulle, et al. 2015; Rambo-Martin et al. 2018). Low to moderate effect common variants that require larger sample sizes to reach genome-wide significance have been proposed to be more likely in these cohorts

(Sailani et al., 2013). The largest GWAS to date of DS-associated AVSD included a cohort of 210 complete AVSD cases with DS (DS+AVSD; diagnosed with full trisomy 21) and 242 controls with DS and structurally normal hearts (DS+NH) (Ramachandran et al., 2015). Several candidate pathways have been reported to be associated with DS AVSD including the folate pathway (Locke et al., 2010), VEGF pathway (Ackerman et al., 2012), and the ciliome (Ripoll et al., 2012).

Figure 1.1. Expression of the FMR1 mRNA and translation into FMRP differs at different sizes of the CGG repeat in the 5' UTR of the FMR1 resulting in different phenotypes. (Adapted from Berman et al., 2014)

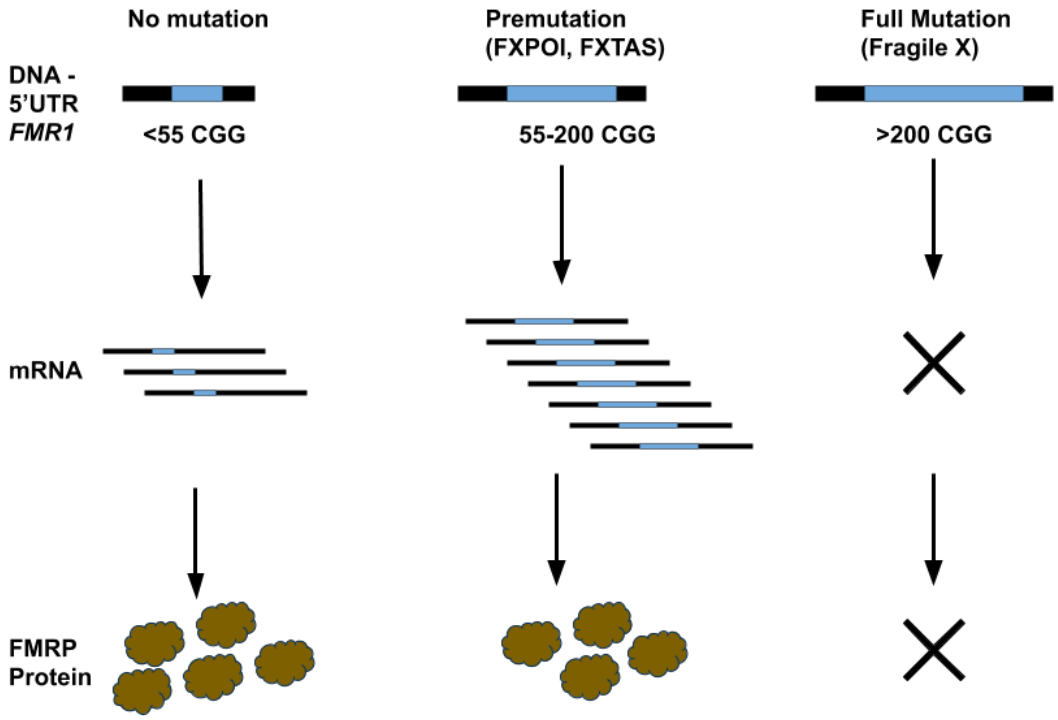
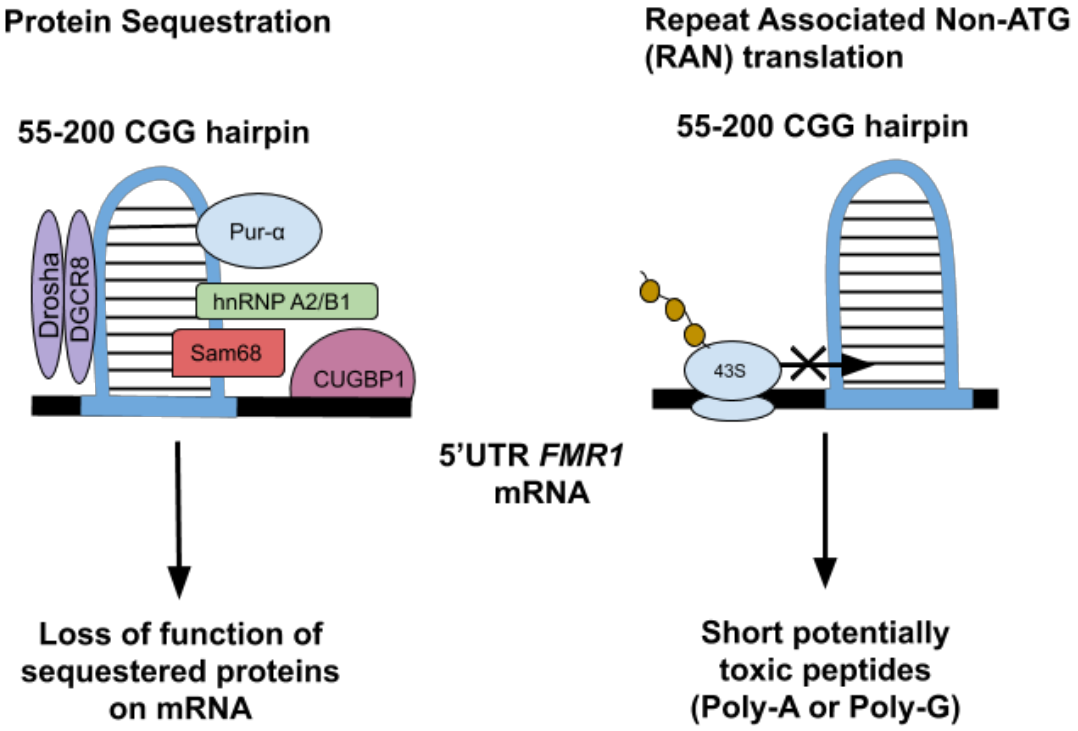


Figure 1.2. Potential mechanisms involved in CGG PM-related pathology. RNA binding proteins can be sequestered in hairpin-like structures of FMR1 mRNA and RAN

translation can form potentially toxic polypeptides when stalled on mRNA structures (Adapted from Berman et al., 2014)



I.V References

Ackerman, C., Locke, A. E., Feingold, E., Reshey, B., Espana, K., Thusberg, J., Mooney, S., Bean, L. J. H., Dooley, K. J., Cua, C. L., Reeves, R. H., Sherman, S. L., & Maslen, C. L. (2012). An excess of deleterious variants in VEGF-A pathway genes in Down-syndrome-associated atrioventricular septal defects. *American Journal of Human Genetics*, *91*(4), 646–659. https://doi.org/10.1016/j.ajhg.2012.08.017

Al Turki, S., Manickaraj, A. K., Mercer, C. L., Gerety, S. S., Hitz, M.-P., Lindsay, S., D'Alessandro, L. C. A., Swaminathan, G. J., Bentham, J., Arndt, A.-K., Louw, J., Low, J., Breckpot, J., Gewillig, M., Thienpont, B., Abdul-Khaliq, H., Harnack, C.,

Hoff, K., Kramer, H.-H., … Hurles, M. E. (2014). Rare variants in NR2F2 cause congenital heart defects in humans. *American Journal of Human Genetics*, *94*(4), 574–585. https://doi.org/10.1016/j.ajhg.2014.03.007

Allen, E. G., Sullivan, A. K., Marcus, M., Small, C., Dominguez, C., Epstein, M. P., Charen, K., He, W., Taylor, K. C., & Sherman, S. L. (2007). Examination of reproductive aging milestones among women who carry the FMR1 premutation. *Human Reproduction (Oxford, England)*, *22*(8), 2142–2152. https://doi.org/10.1093/humrep/dem148

Anagnostis, P., Christou, K., Artzouchaltzi, A.-M., Gkekas, N. K., Kosmidou, N., Siolos, P., Paschou, S. A., Potoupnis, M., Kenanidis, E., Tsiridis, E., Lambrinoudaki, I., Stevenson, J. C., & Goulis, D. G. (2019). Early menopause and premature ovarian insufficiency are associated with increased risk of type 2 diabetes: A systematic review and meta-analysis. *European Journal of Endocrinology*, *180*(1), 41–50. https://doi.org/10.1530/EJE-18-0602

Antonarakis, S. E. (2017). Down syndrome and the complexity of genome dosage imbalance. *Nature Reviews Genetics*, *18*(3), 147–163. https://doi.org/10.1038/nrg.2016.154

Antonarakis, S. E., Skotko, B. G., Rafii, M. S., Strydom, A., Pape, S. E., Bianchi, D. W., Sherman, S. L., & Reeves, R. H. (2020). Down syndrome. *Nature Reviews Disease Primers*, *6*(1), 1–20. https://doi.org/10.1038/s41572-019-0143-7

Ash, P. E. A., Bieniek, K. F., Gendron, T. F., Caulfield, T., Lin, W.-L., DeJesus-Hernandez, M., van Blitterswijk, M. M., Jansen-West, K., Paul, J. W., Rademakers, R., Boylan, K. B., Dickson, D. W., & Petrucelli, L. (2013).

Unconventional Translation of C9ORF72 GGGGCC Expansion Generates Insoluble Polypeptides Specific to c9FTD/ALS. *Neuron*, *77*(4), 639–646. https://doi.org/10.1016/j.neuron.2013.02.004

Aumiller, V., Graebsch, A., Kremmer, E., Niessing, D., & Förstemann, K. (2012). Drosophila Pur-α binds to trinucleotide-repeat containing cellular RNAs and translocates to the early oocyte. *RNA Biology*, *9*(5), 633–643. https://doi.org/10.4161/rna.19760

Berman, R. F., Buijsen, R. A., Usdin, K., Pintado, E., Kooy, F., Pretto, D., Pessah, I. N., Nelson, D. L., Zalewski, Z., Charlet-Bergeurand, N., Willemsen, R., & Hukema, R. K. (2014). Mouse models of the fragile X premutation and fragile X-associated tremor/ataxia syndrome. *Journal of Neurodevelopmental Disorders*, *6*(1), 25. https://doi.org/10.1186/1866-1955-6-25

Bione, S., Benedetti, S., Goegan, M., Menditto, I., Marozzi, A., Ferrari, M., & Toniolo, D. (2006). Skewed X-chromosome inactivation is not associated with premature ovarian failure in a large cohort of Italian patients. *American Journal of Medical Genetics. Part A*, *140*(12), 1349–1351. https://doi.org/10.1002/ajmg.a.31312

Brown, A. L., de Smith, A. J., Gant, V. U., Yang, W., Scheurer, M. E., Walsh, K. M., Chernus, J. M., Kallsen, N. A., Peyton, S. A., Davies, G. E., Ehli, E. A., Winick, N., Heerema, N. A., Carroll, A. J., Borowitz, M. J., Wood, B. L., Carroll, W. L., Raetz, E. A., Feingold, E., … Rabin, K. R. (2019). Inherited genetic susceptibility to acute lymphoblastic leukemia in Down syndrome. *Blood*, *134*(15), 1227–1237. https://doi.org/10.1182/blood.2018890764

Buijsen, R. A. M., Visser, J. A., Kramer, P., Severijnen, E. A. W. F. M., Gearing, M.,

Charlet-Berguerand, N., Sherman, S. L., Berman, R. F., Willemsen, R., & Hukema, R. K. (2016). Presence of inclusions positive for polyglycine containing protein, FMRpolyG, indicates that repeat-associated non-AUG translation plays a role in fragile X-associated primary ovarian insufficiency. *Human Reproduction (Oxford, England)*, *31*(1), 158–168. https://doi.org/10.1093/humrep/dev280

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(Database issue), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Bussani, C., Papi, L., Sestini, R., Baldinotti, F., Bucciantini, S., Bruni, V., & Scarselli, G. (2004). Premature ovarian failure and fragile X premutation: A study on 45 women. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, *112*(2), 189–191. https://doi.org/10.1016/j.ejogrb.2003.06.003

Chalmer, M. A., Esserlind, A.-L., Olesen, J., & Hansen, T. F. (2018). Polygenic risk score: Use in migraine research. *The Journal of Headache and Pain*, *19*(1). https://doi.org/10.1186/s10194-018-0856-0

Chatterjee, N., Shi, J., & García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, *17*(7), 392–406. https://doi.org/10.1038/nrg.2016.27

Cleynen, I., Engchuan, W., Hestand, M. S., Heung, T., Holleman, A. M., Johnston, H.

R., Monfeuga, T., McDonald-McGinn, D. M., Gur, R. E., Morrow, B. E., Swillen, A., Vorstman, J. A. S., Bearden, C. E., Chow, E. W. C., van den Bree, M., Emanuel, B. S., Vermeesch, J. R., Warren, S. T., Owen, M. J., … Bassett, A. S. (2020). Genetic contributors to risk of schizophrenia in the presence of a 22q11.2 deletion. *Molecular Psychiatry*, 1–15. https://doi.org/10.1038/s41380-020-0654-3

Cordell, H. J., Bentham, J., Topf, A., Zelenika, D., Heath, S., Mamasoula, C., Cosgrove, C., Blue, G., Granados-Riveron, J., Setchfield, K., Thornborough, C., Breckpot, J., Soemedi, R., Martin, R., Rahman, T. J., Hall, D., van Engelen, K., Moorman, A. F. M., Zwinderman, A. H., … Keavney, B. D. (2013). Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16. *Nature Genetics*, *45*(7), 822–824. https://doi.org/10.1038/ng.2637

Cordell, H. J., Töpf, A., Mamasoula, C., Postma, A. V., Bentham, J., Zelenika, D., Heath, S., Blue, G., Cosgrove, C., Granados Riveron, J., Darlay, R., Soemedi, R., Wilson, I. J., Ayers, K. L., Rahman, T. J., Hall, D., Mulder, B. J. M., Zwinderman, A. H., van Engelen, K., … Goodship, J. A. (2013). Genome-wide association study identifies loci on 12q24 and 13q32 associated with tetralogy of Fallot. *Human Molecular Genetics*, *22*(7), 1473–1481. https://doi.org/10.1093/hmg/dds552

Craig, B. (2006). Atrioventricular septal defect: From fetus to adult. *Heart*, *92*(12), 1879–1885. https://doi.org/10.1136/hrt.2006.093344

Cronister, A., DiMaio, M., Mahoney, M. J., Donnenfeld, A. E., & Hallam, S. (2005). Fragile X syndrome carrier screening in the prenatal genetic counseling setting.

Genetics in Medicine: Official Journal of the American College of Medical Genetics, *7*(4), 246–250. https://doi.org/10.1097/01.gim.0000159898.90221.d3

Day, F. R., Ruth, K. S., Thompson, D. J., Lunetta, K. L., Pervjakova, N., Chasman, D. I., Stolk, L., Finucane, H. K., Sulem, P., Bulik-Sullivan, B., Esko, T., Johnson, A. D., Elks, C. E., Franceschini, N., He, C., Altmaier, E., Brody, J. A., Franke, L. L., Huffman, J. E., … Murray, A. (2015). Large-scale genomic analyses link reproductive ageing to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nature Genetics*, *47*(11), 1294–1303. https://doi.org/10.1038/ng.3412

Escott-Price, V., Myers, A. J., Huentelman, M., & Hardy, J. (2017). Polygenic Risk Score Analysis of Pathologically Confirmed Alzheimer's Disease. *Annals of Neurology*, *82*(2), 311–314. https://doi.org/10.1002/ana.24999

Ferder, I., Parborell, F., Chiauzzi, V., Gomez, K., Charreau, E., Tesone, M., & Dain, L. (2013). Expression of fragile X mental retardation protein (FMRP) and Fmr1 mRNA during folliculogenesis in the rat. *Reproduction (Cambridge, England)*, *145*. https://doi.org/10.1530/REP-12-0305

Fortuño, C., & Labarta, E. (2014). Genetics of primary ovarian insufficiency: A review. *Journal of Assisted Reproduction and Genetics*, *31*(12), 1573–1585. https://doi.org/10.1007/s10815-014-0342-9

Goswami, D., & Conway, G. S. (2005). Premature ovarian failure. *Human Reproduction Update*, *11*(4), 391–410. https://doi.org/10.1093/humupd/dmi012

Greenway, S. C., Pereira, A. C., Lin, J. C., DePalma, S. R., Israel, S. J., Mesquita, S. M., Ergul, E., Conta, J. R., Korn, J. M., McCarroll, S. A., Gorham, J. M., Gabriel,

S., Altshuler, D. A., de Lourdes Quintanilla-Dieck, M., Artunduaga, M. A., Eavey, R. D., Plenge, R. M., Shadick, N. A., Weinblatt, M. E., … Seidman, C. E. (2009). De Novo Copy Number Variants Identify New Genes and Loci in Isolated, Sporadic Tetralogy of Fallot. *Nature Genetics*, *41*(8), 931–935. https://doi.org/10.1038/ng.415

Hagerman, P. J. (2008). The fragile X prevalence paradox. *Journal of Medical Genetics*, *45*(8), 498–499. https://doi.org/10.1136/jmg.2008.059055

Hantash, F. M., Goos, D. M., Crossley, B., Anderson, B., Zhang, K., Sun, W., & Strom, C. M. (2011). FMR1 premutation carrier frequency in patients undergoing routine population-based carrier screening: Insights into the prevalence of fragile X syndrome, fragile X-associated tremor/ataxia syndrome, and fragile X-associated primary ovarian insufficiency in the United States. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *13*(1), 39–45. https://doi.org/10.1097/GIM.0b013e3181fa9fad

Hartman, R. J., Riehle-Colarusso, T., Lin, A., Frías, J. L., Patel, S. S., Duwe, K., Correa, A., Rasmussen, S. A., & National Birth Defects Prevention Study. (2011). Descriptive study of nonsyndromic atrioventricular septal defects in the National Birth Defects Prevention Study, 1997-2005. *American Journal of Medical Genetics. Part A*, *155A*(3), 555–564. https://doi.org/10.1002/ajmg.a.33874

Hertzberg, L., Vendramini, E., Ganmore, I., Cazzaniga, G., Schmitz, M., Chalker, J., Shiloh, R., Iacobucci, I., Shochat, C., Zeligson, S., Cario, G., Stanulla, M., Strehl, S., Russell, L. J., Harrison, C. J., Bornhauser, B., Yoda, A., Rechavi, G., Bercovich, D., … Izraeli, S. (2010). Down syndrome acute lymphoblastic

leukemia, a highly heterogeneous disease in which aberrant expression of CRLF2 is associated with mutated JAK2: A report from the International BFM Study Group. *Blood*, *115*(5), 1006–1017. https://doi.org/10.1182/blood-2009-08-235408

Hoang, T. T., Goldmuntz, E., Roberts, A. E., Chung, W. K., Kline, J. K., Deanfield, J. E., Giardini, A., Aleman, A., Gelb, B. D., Mac Neal, M., Porter, G. A., Kim, R., Brueckner, M., Lifton, R. P., Edman, S., Woyciechowski, S., Mitchell, L. E., & Agopian, A. J. (2018). The Congenital Heart Disease Genetic Network Study: Cohort description. *PloS One*, *13*(1), e0191319. https://doi.org/10.1371/journal.pone.0191319

Hoffman, G. E., Le, W. W., Entezam, A., Otsuka, N., Tong, Z.-B., Nelson, L., Flaws, J. A., McDonald, J. H., Jafar, S., & Usdin, K. (2012). Ovarian abnormalities in a mouse model of fragile X primary ovarian insufficiency. *The Journal of Histochemistry and Cytochemistry: Official Journal of the Histochemistry Society*, *60*(6), 439–456. https://doi.org/10.1369/0022155412441002

Hunter, J. E., Epstein, M. P., Tinker, S. W., Charen, K. H., & Sherman, S. L. (2008). Fragile X-associated Primary Ovarian Insufficiency: Evidence for Additional Genetic Contributions to Severity. *Genetic Epidemiology*, *32*(6), 553–559. https://doi.org/10.1002/gepi.20329

Jacobsen, B. K., Heuch, I., & Kvåle, G. (2003). Age at Natural Menopause and All-Cause Mortality: A 37-Year Follow-up of 19,731 Norwegian Women. *American Journal of Epidemiology*, *157*(10), 923–929. https://doi.org/10.1093/aje/kwg066

Jin, P., Zarnescu, D. C., Zhang, F., Pearson, C. E., Lucchesi, J. C., Moses, K., &

Warren, S. T. (2003). RNA-mediated neurodegeneration caused by the fragile X premutation rCGG repeats in Drosophila. *Neuron*, *39*(5), 739–747. https://doi.org/10.1016/s0896-6273(03)00533-6

Karmiloff-Smith, A., Al-Janabi, T., D'Souza, H., Groet, J., Massand, E., Mok, K., Startin, C., Fisher, E., Hardy, J., Nizetic, D., Tybulewicz, V., & Strydom, A. (2016). The importance of understanding individual differences in Down syndrome. *F1000Research*, *5*, 389. https://doi.org/10.12688/f1000research.7506.1

Kauppi, K., Westlye, L. T., Tesli, M., Bettella, F., Brandt, C. L., Mattingsdal, M., Ueland, T., Espeseth, T., Agartz, I., Melle, I., Djurovic, S., & Andreassen, O. A. (2015). Polygenic Risk for Schizophrenia Associated With Working Memory-related Prefrontal Brain Activation in Patients With Schizophrenia and Healthy Controls. *Schizophrenia Bulletin*, *41*(3), 736–743. https://doi.org/10.1093/schbul/sbu152

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., & Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, *50*(9), 1219–1224. https://doi.org/10.1038/s41588-018-0183-z

Koon, A. C., & Chan, H. Y. E. (2017). Drosophila melanogaster As a Model Organism to Study RNA Toxicity of Repeat Expansion-Associated Neurodegenerative and Neuromuscular Diseases. *Frontiers in Cellular Neuroscience*, *11*. https://doi.org/10.3389/fncel.2017.00070

Lévesque, S., Dombrowski, C., Morel, M.-L., Rehel, R., Côté, J.-S., Bussières, J., Morgan, K., & Rousseau, F. (2009). Screening and instability of FMR1 alleles in

a prospective sample of 24,449 mother-newborn pairs from the general population. *Clinical Genetics*, *76*(6), 511–523. https://doi.org/10.1111/j.1399-0004.2009.01237.x

Lin, F.-J., You, L.-R., Yu, C.-T., Hsu, W.-H., Tsai, M.-J., & Tsai, S. Y. (2012). Endocardial Cushion Morphogenesis and Coronary Vessel Development Require Chicken Ovalbumin Upstream Promoter-Transcription Factor II. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *32*(11), e135–e146. https://doi.org/10.1161/ATVBAHA.112.300255

Locke, A. E., Dooley, K. J., Tinker, S. W., Cheong, S. Y., Feingold, E., Allen, E. G., Freeman, S. B., Torfs, C. P., Cua, C. L., Epstein, M. P., Wu, M. C., Lin, X., Capone, G., Sherman, S. L., & Bean, L. J. H. (2010). Variation in folate pathway genes contributes to risk of congenital heart defects among individuals with Down syndrome. *Genetic Epidemiology*, *34*(6), 613–623. https://doi.org/10.1002/gepi.20518

Lu, C., Lin, L., Tan, H., Wu, H., Sherman, S. L., Gao, F., Jin, P., & Chen, D. (2012). Fragile X premutation RNA is sufficient to cause primary ovarian insufficiency in mice. *Human Molecular Genetics*, *21*(23), 5039–5047. https://doi.org/10.1093/hmg/dds348

Luo, Y., Hitz, B. C., Gabdank, I., Hilton, J. A., Kagda, M. S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., Baymuradov, U. K., Graham, K., Litton, C., Miyasato, S. R., Strattan, J. S., Jolanki, O., Lee, J.-W., Tanaka, F. Y., Adenekan, P., … Cherry, J. M. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research*, *48*(D1), D882–D889.

https://doi.org/10.1093/nar/gkz1062

Mai, C. T., Isenburg, J., Langlois, P. H., Alverson, C. J., Gilboa, S. M., Rickard, R., Canfield, M. A., Anjohrin, S. B., Lupo, P. J., Jackson, D. R., Stallings, E. B., Scheuerle, A. E., Kirby, R. S., & National Birth Defects Prevention Network. (2015). Population-based birth defects data in the United States, 2008 to 2012: Presentation of state-specific data and descriptive brief on variability of prevalence. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, *103*(11), 972–993. https://doi.org/10.1002/bdra.23461

Marozzi, A., Vegetti, W., Manfredini, E., Tibiletti, M. G., Testa, G., Crosignani, P. G., Ginelli, E., Meneveri, R., & Dalprà, L. (2000). Association between idiopathic premature ovarian failure and fragile X premutation. *Human Reproduction*, *15*(1), 197–202. https://doi.org/10.1093/humrep/15.1.197

Mori, K., Weng, S.-M., Arzberger, T., May, S., Rentzsch, K., Kremmer, E., Schmid, B., Kretzschmar, H. A., Cruts, M., Broeckhoven, C. V., Haass, C., & Edbauer, D. (2013). The C9orf72 GGGGCC Repeat Is Translated into Aggregating Dipeptide-Repeat Proteins in FTLD/ALS. *Science*, *339*(6125), 1335–1338. https://doi.org/10.1126/science.1232927

Moutsianas, L., Agarwala, V., Fuchsberger, C., Flannick, J., Rivas, M. A., Gaulton, K. J., Albers, P. K., Consortium, G., McVean, G., Boehnke, M., Altshuler, D., & McCarthy, M. I. (2015). The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. *PLOS Genetics*, *11*(4), e1005165. https://doi.org/10.1371/journal.pgen.1005165

Muka, T., Oliver-Williams, C., Kunutsor, S., Laven, J. S. E., Fauser, B. C. J. M., Chowdhury, R., Kavousi, M., & Franco, O. H. (2016). Association of Age at Onset of Menopause and Time Since Onset of Menopause With Cardiovascular Outcomes, Intermediate Vascular Traits, and All-Cause Mortality: A Systematic Review and Meta-analysis. *JAMA Cardiology*, *1*(7), 767–776. https://doi.org/10.1001/jamacardio.2016.2415

Murray, A. (2000). Premature Ovarian Failure and the FMR1 Gene. *Seminars in Reproductive Medicine*, *18*(01), 059–066. https://doi.org/10.1055/s-2000-13476

Nelson, L. M. (2009). Clinical practice. Primary ovarian insufficiency. *The New England Journal of Medicine*, *360*(6), 606–614. https://doi.org/10.1056/NEJMcp0808697

Pereira, F. A., Qiu, Y., Zhou, G., Tsai, M.-J., & Tsai, S. Y. (1999). The orphan nuclear receptor COUP-TFII is required for angiogenesis and heart development. *Genes & Development*, *13*(8), 1037–1049.

Perry, J. R. B., Hsu, Y.-H., Chasman, D. I., Johnson, A. D., Elks, C., Albrecht, E., Andrulis, I. L., Beesley, J., Berenson, G. S., Bergmann, S., Bojesen, S. E., Bolla, M. K., Brown, J., Buring, J. E., Campbell, H., Chang-Claude, J., Chenevix-Trench, G., Corre, T., Couch, F. J., … Murray, A. (2014). DNA mismatch repair gene MSH6 implicated in determining age at natural menopause. *Human Molecular Genetics*, *23*(9), 2490–2497. https://doi.org/10.1093/hmg/ddt620

Pierpont, M. E., Basson, C. T., Benson, D. W., Gelb, B. D., Giglia, T. M., Goldmuntz, E., McGee, G., Sable, C. A., Srivastava, D., Webb, C. L., & American Heart Association Congenital Cardiac Defects Committee, Council on Cardiovascular Disease in the Young. (2007). Genetic basis for congenital heart defects: Current

knowledge: a scientific statement from the American Heart Association Congenital Cardiac Defects Committee, Council on Cardiovascular Disease in the Young: endorsed by the American Academy of Pediatrics. *Circulation*, *115*(23), 3015–3038. https://doi.org/10.1161/CIRCULATIONAHA.106.183056

Ramachandran, D., Mulle, J. G., Locke, A. E., Bean, L. J. H., Rosser, T. C., Bose, P., Dooley, K. J., Cua, C. L., Capone, G. T., Reeves, R. H., Maslen, C. L., Cutler, D. J., Sherman, S. L., & Zwick, M. E. (2015). Contribution of copy-number variation to Down syndrome-associated atrioventricular septal defects. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *17*(7), 554–560. https://doi.org/10.1038/gim.2014.144

Reddy, K., & Pearson, C. E. (2013). RAN Translation: Fragile X in the Running. *Neuron*, *78*(3), 405–408. https://doi.org/10.1016/j.neuron.2013.04.034

Reller, M. D., Strickland, M. J., Riehle-Colarusso, T., Mahle, W. T., & Correa, A. (2008). Prevalence of Congenital Heart Defects in Metropolitan Atlanta, 1998–2005. *The Journal of Pediatrics*, *153*(6), 807–813. https://doi.org/10.1016/j.jpeds.2008.05.059

Ripoll, C., Rivals, I., Ait Yahya-Graison, E., Dauphinot, L., Paly, E., Mircher, C., Ravel, A., Grattau, Y., Bléhaut, H., Mégarbane, A., Dembour, G., de Fréminville, B., Touraine, R., Créau, N., Potier, M. C., & Delabar, J. M. (2012). Molecular Signatures of Cardiac Defects in Down Syndrome Lymphoblastoid Cell Lines Suggest Altered Ciliome and Hedgehog Pathways. *PLoS ONE*, *7*(8). https://doi.org/10.1371/journal.pone.0041616

Robinson, S. W., Morris, C. D., Goldmuntz, E., Reller, M. D., Jones, M. A., Steiner, R.

D., & Maslen, C. L. (2003). Missense Mutations in CRELD1 Are Associated with Cardiac Atrioventricular Septal Defects. *American Journal of Human Genetics*, *72*(4), 1047–1052.

Rodriguez-Revenga, L., Madrigal, I., Badenas, C., Xunclà, M., Jiménez, L., & Milà, M. (2009). Premature ovarian failure and fragile X female premutation carriers: No evidence for a skewed X-chromosome inactivation pattern. *Menopause (New York, N.Y.)*, *16*(5), 944–949. https://doi.org/10.1097/gme.0b013e3181a06a37

Rudnicka, E., Kruszewska, J., Klicka, K., Kowalczyk, J., Grymowicz, M., Skórska, J., Pięta, W., & Smolarczyk, R. (2018). Premature ovarian insufficiency – aetiopathology, epidemiology, and diagnostic evaluation. *Przegląd Menopauzalny = Menopause Review*, *17*(3), 105–108. https://doi.org/10.5114/pm.2018.78550

Sailani, M. R., Makrythanasis, P., Valsesia, A., Santoni, F. A., Deutsch, S., Popadin, K., Borel, C., Migliavacca, E., Sharp, A. J., Duriaux Sail, G., Falconnet, E., Rabionet, K., Serra-Juhé, C., Vicari, S., Laux, D., Grattau, Y., Dembour, G., Megarbane, A., Touraine, R., … Antonarakis, S. E. (2013). The complex SNP and CNV genetic architecture of the increased risk of congenital heart defects in Down syndrome. *Genome Research*, *23*(9), 1410–1421. https://doi.org/10.1101/gr.147991.112

Sarısoy, Ö., Ayabakan, C., Tokel, K., Özkan, M., Türköz, R., & Aşlamacı, S. (2018). Long-term outcomes in patients who underwent surgical correction for atrioventricular septal defect. *Anatolian Journal of Cardiology*, *20*(4), 229–234. https://doi.org/10.14744/AnatolJCardiol.2018.39660

Sellier, C., Freyermuth, F., Tabet, R., Tran, T., He, F., Ruffenach, F., Alunni, V., Moine,

H., Thibault, C., Page, A., Tassone, F., Willemsen, R., Disney, M. D., Hagerman, P. J., Todd, P. K., & Charlet-Berguerand, N. (2013). Sequestration of DROSHA and DGCR8 by expanded CGG RNA repeats alters microRNA processing in fragile X-associated tremor/ataxia syndrome. *Cell Reports*, *3*(3), 869–880. https://doi.org/10.1016/j.celrep.2013.02.004

Sellier, C., Rau, F., Liu, Y., Tassone, F., Hukema, R. K., Gattoni, R., Schneider, A., Richard, S., Willemsen, R., Elliott, D. J., Hagerman, P. J., & Charlet-Berguerand, N. (2010). Sam68 sequestration and partial loss of function are associated with splicing alterations in FXTAS patients. *The EMBO Journal*, *29*(7), 1248–1261. https://doi.org/10.1038/emboj.2010.21

Sherman, S. L. (2000). Premature ovarian failure in the fragile X syndrome. *American Journal of Medical Genetics*, *97*(3), 189–194. https://doi.org/10.1002/1096-8628(200023)97:3<189::AID-AJMG1036>3.0.CO;2-J

Sherman, Stephanie L, Curnow, E. C., Easley, C. A., Jin, P., Hukema, R. K., Tejada, M. I., Willemsen, R., & Usdin, K. (2014). Use of model systems to understand the etiology of fragile X-associated primary ovarian insufficiency (FXPOI). *Journal of Neurodevelopmental Disorders*, *6*(1), 26. https://doi.org/10.1186/1866-1955-6-26

Shuster, L. T., Rhodes, D. J., Gostout, B. S., Grossardt, B. R., & Rocca, W. A. (2010). Premature menopause or early menopause: Long-term health consequences. *Maturitas*, *65*(2), 161–166. https://doi.org/10.1016/j.maturitas.2009.08.003

Soemedi, R., Wilson, I. J., Bentham, J., Darlay, R., Töpf, A., Zelenika, D., Cosgrove, C., Setchfield, K., Thornborough, C., Granados-Riveron, J., Blue, G. M., Breckpot, J., Hellens, S., Zwolinkski, S., Glen, E., Mamasoula, C., Rahman, T. J., Hall, D.,

Rauch, A., … Keavney, B. D. (2012). Contribution of Global Rare Copy-Number Variants to the Risk of Sporadic Congenital Heart Disease. *American Journal of Human Genetics*, *91*(3), 489–501. https://doi.org/10.1016/j.ajhg.2012.08.003

Sofola, O. A., Jin, P., Qin, Y., Duan, R., Liu, H., de Haro, M., Nelson, D. L., & Botas, J. (2007). RNA-binding proteins hnRNP A2/B1 and CUGBP1 suppress fragile X CGG premutation repeat-induced neurodegeneration in a Drosophila model of FXTAS. *Neuron*, *55*(4), 565–571. https://doi.org/10.1016/j.neuron.2007.07.021

Song, F. J., Barton, P., Sleightholme, V., Yao, G. L., & Fry-Smith, A. (2003). Screening for fragile X syndrome: A literature review and modelling study. *Health Technology Assessment (Winchester, England)*, *7*(16), 1–106. https://doi.org/10.3310/hta7160

Spath, M. A., Feuth, T. B., Smits, A. P. T., Yntema, H. G., Braat, D. D. M., Thomas, C. M. G., van Kessel, A. G., Sherman, S. L., & Allen, E. G. (2011). Predictors and risk model development for menopausal age in fragile x premutation carriers. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *13*(7), 643–650. https://doi.org/10.1097/GIM.0b013e31821705e5

Spath, M. A., Nillesen, W. N., Smits, A. P. T., Feuth, T. B., Braat, D. D. M., van Kessel, A. G., & Yntema, H. G. (2010). X chromosome inactivation does not define the development of premature ovarian failure in fragile X premutation carriers. *American Journal of Medical Genetics. Part A*, *152A*(2), 387–393. https://doi.org/10.1002/ajmg.a.33243

Stolk, L., Perry, J. R., Chasman, D. I., He, C., Mangino, M., Sulem, P., Barbalic, M., Broer, L., Byrne, E. M., Ernst, F., Esko, T., Franceschini, N., Gudbjartsson, D. F.,

Hottenga, J.-J., Kraft, P., McArdle, P. F., Porcu, E., Shin, S.-Y., Smith, A. V., … Lunetta, K. L. (2012). Meta-analyses identify 13 novel loci associated with age at menopause and highlights DNA repair and immune pathways. *Nature Genetics*, *44*(3), 260–268. https://doi.org/10.1038/ng.1051

Tejada, M.-I., García-Alegría, E., Bilbao, A., Martínez-Bouzas, C., Beristain, E., Poch, M., Ramos-Arroyo, M. A., López, B., Fernandez Carvajal, I., Ribate, M.-P., & Ramos, F. (2008). Analysis of the molecular parameters that could predict the risk of manifesting premature ovarian failure in female premutation carriers of fragile X syndrome. *Menopause (New York, N.Y.)*, *15*(5), 945–949. https://doi.org/10.1097/gme.0b013e3181647762

Todd, P. K., Oh, S. Y., Krans, A., He, F., Sellier, C., Frazer, M., Renoux, A. J., Chen, K., Scaglione, K. M., Basrur, V., Elenitoba-Johnson, K., Vonsattel, J. P., Louis, E. D., Sutton, M. A., Taylor, J. P., Mills, R. E., Charlet-Berguerand, N., & Paulson, H. L. (2013). CGG repeat-associated translation mediates neurodegeneration in fragile X tremor ataxia syndrome. *Neuron*, *78*(3), 440–455. https://doi.org/10.1016/j.neuron.2013.03.026

Torkamani, A., Wineinger, N. E., & Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, *19*(9), 581–590. https://doi.org/10.1038/s41576-018-0018-x

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Vujovic, S. (2009). Aetiology of premature ovarian failure. *Menopause International*, *15*(2), 72–75. https://doi.org/10.1258/mi.2009.009020

Webber, L., Davies, M., Anderson, R., Bartlett, J., Braat, D., Cartwright, B., Cifkova, R., de Muinck Keizer-Schrama, S., Hogervorst, E., Janse, F., Liao, L., Vlaisavljevic, V., Zillikens, C., & Vermeulen, N. (2016). ESHRE Guideline: Management of women with premature ovarian insufficiency. *Human Reproduction*, *31*(5), 926–937. https://doi.org/10.1093/humrep/dew027

Yang, L., Duan, R., Chen, D., Wang, J., Chen, D., & Jin, P. (2007). Fragile X mental retardation protein modulates the fate of germline stem cells in Drosophila. *Human Molecular Genetics*, *16*(15), 1814–1820. https://doi.org/10.1093/hmg/ddm129

Yang, Y., Xu, S., Xia, L., Wang, J., Wen, S., Jin, P., & Chen, D. (2009). The bantam microRNA is associated with drosophila fragile X mental retardation protein and regulates the fate of germline stem cells. *PLoS Genetics*, *5*(4), e1000444. https://doi.org/10.1371/journal.pgen.1000444

Zaidi, S., Choi, M., Wakimoto, H., Ma, L., Jiang, J., Overton, J. D., Romano-Adesman, A., Bjornson, R. D., Breitbart, R. E., Brown, K. K., Carriero, N. J., Cheung, Y. H., Deanfield, J., DePalma, S., Fakhro, K. A., Glessner, J., Hakonarson, H., Italia, M. J., Kaltman, J. R., … Lifton, R. P. (2013). De novo mutations in histone-modifying genes in congenital heart disease. *Nature*, *498*(7453), 220–223. https://doi.org/10.1038/nature12141

Zu, T., Gibbens, B., Doty, N. S., Gomes-Pereira, M., Huguet, A., Stone, M. D., Margolis, J., Peterson, M., Markowski, T. W., Ingram, M. A. C., Nan, Z., Forster, C., Low,

W. C., Schoser, B., Somia, N. V., Clark, H. B., Schmechel, S., Bitterman, P. B., Gourdon, G., … Ranum, L. P. W. (2011). Non-ATG–initiated translation directed by microsatellite expansions. *Proceedings of the National Academy of Sciences*, *108*(1), 260–265. https://doi.org/10.1073/pnas.1013343108

II. Identifying modifying genes to explain the variation in severity of fragile X-associated primary ovarian insufficiency

Cristina E. Trevino, J. Christopher Rounds, Emily G. Allen, Peng Jin, H. Richard Johnston, Dave J. Cutler, Mike E. Zwick, Stephen T. Warren, Stephanie L. Sherman

Abstract

Primary ovarian insufficiency (POI) affects 1% of women in the general population and is characterized by symptoms of early menopause and infertility. If untreated, it leads to an increased risk of disorders related to early estrogen deficiency. Those who carry an *FMR1* premutation, a CGG repeat expansion in the range of 55-200 repeats in the 5'UTR of the X-linked *FMR1* gene, are at a 20-fold increased risk for fragile-X associated POI (FXPOI). Although the risk for FXPOI is significant, not all carriers experience the disorder, and the range of severity is broad. We hypothesize that genetic variants, in addition to the premutation repeat number, modify the age of onset of FXPOI. To test this hypothesis, we compared WGS data from women with a premutation who experienced FXPOI before age 35 (cases; n=63) to those with a premutation who experienced age at menopause after age 50 (controls; n=51). Genetic variants were prioritized using a well-established pipeline, generating candidates through gene burden testing of rare variants. The top ranked genes ($p<0.001$) were then functionally screened using the *Drosophila* premutation model to test whether ovarian function was disrupted. Three candidate genes came to the forefront when knocked down: *SUMO1* and *KRR1*, which appeared to play a synergistic role with the premutation, and *PDHA2,* which appear to have an additive effect. We suggest that these new candidate genes, once further studied in mammalian systems, will provide insight into the etiology of FXPOI and potentially point to targeted treatments.

Introduction

Fragile X-associated primary ovarian insufficiency (FXPOI) is one of the disorders associated with the fragile X premutation repeat expansion (55-200 unmethylated CGG repeats) located in the 5'UTR of the *FMR1* gene. It is characterized by amenorrhea for at least 4 months before the age of 40 and altered hormone levels, specifically high follicle stimulating hormone (FSH) and low anti-Mullerian hormone (AMH), that are associated with diminished ovarian reserve (Welt, Smith, and Taylor 2004; Nelson 2009). Women with a premutation (PM) are at a 20% risk of primary ovarian insufficiency (POI) compared to 1% of women in the general population (S. L. Sherman 2000; Nelson 2009), a 20-fold increased risk. However, not all women with a PM suffer from POI. Identification of risk factors for FXPOI, and POI in general, can help predict the potential of a shortened reproductive life span and provide possible interventions to help achieve family building plans and reduce the risk of untreated early estrogen deficiency.

Genetic factors that have been investigated in women with a PM to explain the incomplete penetrance of POI include PM CGG repeat length, skewing of X-chromosome inactivation (XCI), and genetic background. In women with a PM, repeat size is nonlinearly associated with FXPOI, with the greatest risk incurred at 80-100 repeats rather than with the largest PM alleles (Allen et al. 2007; Sullivan et al. 2005; Spath et al. 2011; Ennis, Ward, and Murray 2006). To date, skewed XCI nor the increased percentage of active X chromosomes harboring a PM have not been associated with a higher risk for FXPOI (Tejada et al. 2008; Bione et al. 2006; Rodriguez-Revenga et al. 2009; Spath et al. 2010). Two studies have provided indirect

evidence for modifying genes being involved in explaining the risk of FXPOI. First, evidence for an additive genetic component, adjusting for repeat size, was identified in a large sample of PM carriers and noncarriers (Hunter et al. 2008). Second, the average age of menopause among first degree relatives of PM carriers was found to be associated with the risk for FXPOI (Spath et al. 2011). These findings suggest a significant polygenic component involved in the genetic architecture of onset of FXPOI. Indeed, evidence for polygenes comes from studies in the general population of women and natural age at menopause (Day et al. 2015; Stolk et al. 2012; Perry et al. 2014). For example, the large GWAS presented by Day et al. (2015) identified over 50 common variants associated with natural age of menopause (Day et al. 2015; Stolk et al. 2012; Perry et al. 2014).

Environmental risk factors have been associated with idiopathic POI, including phthalates, bisphenol A, pesticides and tobacco use (Vabre et al. 2017). The primary environmental risk factor that has been identified for risk of FXPOI is smoking (Allen et al. 2007; Spath et al. 2011). The role of active smoking in decreasing natural age of menopause is also found among all women (Cooper, Sandler, and Bohlig 1999) and was found to have the same additive effect on age of onset of FXPOI (Allen et al. 2007). Use of oral contraceptives or hormone replacement treatment (HRT) does not appear to increase the risk of FXPOI, but can mask POI-associated symptoms, thus, complicating the diagnosis of FXPOI (Hunter et al. 2008).

Increased repeat size within the PM range is associated with increased transcription of *FMR1* mRNA, although FMRP levels are the same or reduced (Primerano et al. 2002; Tassone, Hagerman, Taylor, Gane, et al. 2000; Tassone,

Hagerman, Taylor, Mills, et al. 2000; Tassone and Hagerman 2003). Unlike the full mutation (>200 methylated CGG repeats) where the *FMR1* gene is transcriptionally silenced leading to fragile X syndrome, the protein encoded by *FMR1* (FMRP), is still produced by a PM allele (Pieretti et al. 1991; Kenneson et al. 2001). Much has been learned about potential PM-associated molecular mechanisms from fragile X-associated tremor/ataxia syndrome (FXTAS), the other well-established PM-associated disorder. For this neurodegenerative disorder, the toxic effect of the PM is found to be related to the long PM repeat track in the *FMR1* mRNA. This repeat track has the potential to form secondary structures such as hairpins that cause subsequent altered processes (Handa, Saha, and Usdin 2003). Evidence for at least two mechanisms have been identified. First, increased *FMR1* mRNA containing hairpin loops and other structures formed within the PM-size CGG repeats has been shown to sequester specific RNA binding proteins, altering their normal functions (Sellier et al. 2010; 2013; Handa, Saha, and Usdin 2003). Second, repeat-associated non-ATG (RAN) translation, caused by translation machinery becoming stalled on a structure like the hairpins that form in CGG mRNA, produce small potentially toxic polypeptides, in this case alanine or glutamine polymers (Todd et al. 2013; Sellier et al. 2017). Presumably, these two mechanisms also play a role in FXPOI. Murine PM model systems have begun to unravel their importance. All PM models showed traits associated with reduced ovarian function. Overall, it appears that the original follicular pool is not disturbed, but that there is an increased rate of atresia/apoptosis. Evidence from these models indicate that the ovarian phenotype is due to the toxic effect of the *FMR1* mRNA (Reviewed in Sherman et al. 2014).

The *Drosophila* PM model system has proven to be an effective model for screening of neuronal phenotypes associated with FXTAS (Jin et al. 2003). For example, this model clearly showed that specific CGG RNA binding proteins, including *hnRNP A2/A1*, *CUGPB1*, and *Pur-alpha,* alter neuronal function via sequestration of these proteins (Sofola et al. 2007; Aumiller et al. 2012). Germline stem cells in *Drosophila* ovaries have also been evaluated with respect to FMRP (L. Yang et al. 2007; Y. Yang et al. 2009). We took advantage of these established models and used them as a way to screen our genetic results from our whole genome sequencing (WGS) human studies. We used altered fecundity levels as a reporter of ovarian dysfunction to screen for multiple candidate genes, as this system has been established in studies of metabolic pathways (Daenzer et al. 2012; Armstrong 2020).

The goal of this study was to identify potential modifiers of FXPOI. We compared WGS data from women with a PM who experienced FXPOI before age 35 (cases; n=63) to those with a PM who experienced age at menopause after age 50 (controls; n=51) using gene-set analyses. Both an untargeted approach was used as well as a candidate gene approach, focusing of RNA-binding proteins known to bind to the *FMR1* PM mRNA. Highly ranked genes were then screened using *Drosophila* as a whole-organism functional assay.

Subjects and Methods

Participants

Participants and samples were identified through three primary sources, all coordinated through the National Fragile X Center at Emory University. The majority were recruited through the Center's infrastructure that identifies families with a history of fragile X-associated disorders through national and international sources, including the Fragile X Clinics, general genetics clinics, Fragile X Clinic and Research Consortium, fragile X family conferences, fragile X listservs, and parent support groups. Once a family contact is identified, their other family members are screened for the fragile X mutation. Second, the Fragile X Research Registry, a national collaborative effort, used their resources to identify possible participants who then directly contacted the Emory team for consenting and recruiting. Once a participant was consented, a blood or saliva sample was collected and each completed a reproductive and health history questionnaire. Data included general demographics (e.g., age at interview, date of birth, race/ethnicity), lifestyle factors that might affect overall health (e.g., smoking, body mass index), and reproductive history (e.g., menstrual history, reason for cessation of menses, pregnancy history). Protocols and consent forms were approved by Emory University Institutional Review Board, and informed consent was obtained from all participants. Finally, de-identified samples from other collaborators were also included if they met case/control definitions and had appropriate consent for sample sharing.

For this study, all cases and controls included women who carried a premutation, defined as an *FMR1* repeat allele with 55-199 unmethylated CGG repeats. Cases were further defined as those who had cessation of menses for one year prior to age 35 due to FXPOI. Controls were premutation carriers who went through natural menopause, or

cessation of menses for one year, after age 50 and who had no indications of infertility during her reproductive lifespan. We excluded women whose age at menopause could have been affected by FXPOI-unrelated medical conditions, including chemotherapy or radiation therapy, missing ovaries or ovarian or tubal surgery or an eating disorder.

Laboratory Methods

DNA extraction: DNA was extracted from biological samples using Qiagen Qiamp DNA Blood Mini Kit, Gentra Puregene extraction kit, or prepIT-L2P protocol from Oragene.

FMR1 CGG repeat numbers: Premutation status was determined by a fluorescent sequencer method (ref-27). For females with only one allele, a second polymerase chain reaction (PCR) protocol was used (ref-28). The PCRs for FRAXA consisted of 1X PCR Buffer (Gibco/BRL), 10% dimethyl sulfoxide (DMSO), 370 µM deazaG, 500 µM d(ACT), 0.3 µM each primer, 15 ng T4 gene 32, and 1.05 U Roche Expand Long Taq . Primers for the FMR1 gene were C: 5′ GCTCAGCTCCGTTTCGGTTTCACTTCCGGT3′ , and F:5′AGCCCCGCACTTCCACCAGCTCCTCCA3′ (ref-29).

Bioinformatic Analysis

Whole genome sequencing (WGS) was performed on 68 cases and 55 controls for this preliminary analysis by HudsonAlpha (Huntsville, AL). FASTQ files from paired-end WGS reads were mapped and variants were called with PEMapper and PECaller, respectively (Johnston et al. 2017). Variants were annotated using Bystro (Kotlar et al. 2018) (http://bystro.io). A total of 13,808,870 single nucleotide variants (SNV) were

detected by WGS across 68 cases and 55 controls. Mean coverage depth ± standard deviation (sd) of WGS was 30.783 ± 7.090 for samples and mean transition/transversion ratio ± sd was 2.056 ± 0.008.

Sample failures were addressed by removing any individuals missing > 1% genotypes or failing PLINK1.9's sex check (based on F statistics for X chromosome heterozygosity, which were also used to impute sex on individuals missing sex data) (Weir and Cockerham 1984; Chang et al. 2015, Purcell and Chang, n.d.). These filters identified no samples for exclusion. Variant filters included removing SNVs with missingness > 10% and those failing the exact test for HWE at a p-value < 10^{-6}.

We then performed principal component analysis (PCA), using PLINK1.9 (Chang et al. 2015) to identify population stratification. We used common SNPs (MAF > 0.05) and pruned SNPs in linkage disequilibrium with an $r_2 > 0.2$, stepping along five SNPs at a time within 50kb windows. Through three rounds of PCA we identified a total of 9 outlier samples for removal. Following QC, the dataset contained 114 samples (63 cases, 51 controls) and 13,663,751 SNVs for analysis and none of the PCs were significant in the model so they were not included as covariates.

Common variant analysis

Common variants defined as those with MAF > 0.05 from gnomAD genomes data. Logistic regression was performed with PLINK 1.9 of the common variants using age and repeat and repeat size squared as covariates.

Rare variant analysis

Rare variants were defined at minor allele frequency less than 0.05 from gnomAD genomes data. Variants in which the reference allele was the minor allele were excluded. SKAT-O, SKAT and burden testing was done using the SKAT package in R (*R: A Language and Environment for Statistical Computing* 2017) and the genes with the lowest p-value were evaluated as candidate genes. No genes reached Bonferroni correction in any of the rare variant tests and were evaluated from a ranking perspective given the small sample size. For candidate gene ranking, genes with p-values < 0.001 were checked for having fly orthologs, any literature references to ovarian phenotypes, and ovarian expression using GTEx.

Polygenic Risk Score

Our target dataset used to calculate polygenic risk scores (PRS) included 63 PM cases with early AOM and 51 PM controls with who began menopause after age 55. The same standard QC measures described above were used prior to analyzing this dataset as well as removing the major histocompatibility complex region (Chr6: 25-34 Mb, hg19), a region of extended high linkage disequilibrium that can overly influence PRS results. The final target dataset included 724,760 total variants. The discovery dataset used to calculate polygenic risk scores (PRS) was a large GWAS study that included 69,360 women who experienced natural age at menopause (Day et al. 2015). PRSice-2 software was used (Choi and O'Reilly 2019) to measure the proportion of variance in FXPOI case-control status explained (measured by Nagelkerke's R_2) by the PRS using

different p-value thresholds derived from the GWAS study of Day et al 2015. The model also included repeat and repeat size squared as covariates.

In order to obtain an independent set of SNPs for scoring PRSice performs clumping on the discovery dataset (clumping parameters: 500kb window, r_2 threshold 0.10). These clumped SNPs are used to generate PRS, calculated by the following equation:

$$PRS_j = \sum_i \frac{\beta_i \times EA_{ij}}{N_j}$$

in which the subscript *i* denotes a specific SNP contributing to the PRS, the subscript *j* denotes a particular individual in the target dataset, *β* is the estimated effect from the discovery GWAS (e.g., the natural logarithm of the odds ratio), *EA* is the number of effective alleles possessed by the target individual (0,1 or 2 for a disomic chromosome), and *N* is the total number of alleles considered for scoring.

Generation of a stable line expressing 90 CGG in the *Drosophila* germline

Drosophila with the PM repeat (90 CGG repeats) inserted on chromosome 2 were generated as described in P. Jin et al. 2003 obtained from Dr. Peng Jin's lab. Progeny of PM repeat flies and a germline-expressing *nanos>Gal4* line (Bloomington Stock #4442) were generated and crossed to a *Sp*/*CyO* stock to allow for capture of premutation, *nanos>Gal4* recombinant chromosomes. Recombinant males were confirmed through PCR genotyping. Then, *nanos>Gal4,90CGG*/*Sp* males were crossed with a *Sp*/*CyO*, *tubulin>Gal80* stock to obtain a stable, balanced line *nanos>Gal4*,

90CGG/*CyO,tubulin>Gal80*. Based on candidate gene selection guided by the human WGS rare variant analysis and from previously identified candidate RNA-binding proteins, *Drosophila* TRiP lines expressing RNAi constructs against candidate genes were obtained from the Bloomington Drosophila Stock Center (Table 2.1) . Stocks carrying these RNAi constructs were then crossed with both germline-expressing *nanos>Gal4* alone and the *nanos>Gal4*, *90CGG* premutation recombinant for fecundity experiments.

Fecundity Testing

All *Drosophila* stocks were raised at 25˚C on standard media. Fecundity was tested using cages with yeast-supplemented grape juice agar egg-laying plates with 5 females collected within 24 hours after eclosion. Plates were changed out at increments of 24 hours for 10 days. At least three replicates were done per genotype for the initial screen. Control stocks and the stable 90 CGG premutation alone were both crossed with RNAi background stocks (attP docking site lines, Bloomington Stocks #36303 or #36304) to establish baseline fecundity. Candidate gene knockdown lines were crossed with either *nanos>Gal4* alone or *nanos>Gal4,90CGG*/*CyO,tubulin>Gal80*, and non-*CyO* progeny were tested for fecundity. Each candidate gene knockdown was compared to the baseline fecundity values established with controls crossed with Bloomington TRiP background lines (Bloomington Stocks # 36303 or # 36304). To further examine top candidates from the initial screen, a followup screen was conducted with at least 10 replicates to increase sample size to ensure robust results. Plates were imaged using a Nikon D3400 DSLR and processed and counted using Fiji (Schindelin et al. 2015)

image software. Critically, egg counting was conducted blind to the genotype of the fly. Specifically, plates were coded prior to imaging by a first experimenter and scored after imaging by a second experimenter. The outcome fecundity measure analyzed in subsequent regression models was the 10-day total egg count per cage. Experiments where one or more flies died during the course of the 10 days were excluded from the analysis. Due to the overdispersion observed in the data, a quasipoisson regression was used to test for altered fecundity compared with controls based on the main predictors of presence of the candidate gene knockdown, presence of the 90 CGG repeat, and the interaction term between those two genotypes.

Results

In the cohort of 114 PM women, WGS from 63 cases and 51 controls were analyzed. The mean age of menopause was 29.7 for cases and 51.6 for controls. The average PM repeat size was not significantly different between cases and controls (88.3 repeats for cases and 89.5 repeats for controls; $p>0.10$), although the SD was significantly larger for cases (Figure 2.1). The cohort consisted of women who identified as Caucasian. The mean repeat size is not statistically different between cases and controls but cases more often have alleles in the mid-range of 80-100, reflecting the high risk repeat range.

Genome wide association study of common variants

3,055,728 single nucleotide polymorphisms with minor allele frequency (MAF) > 0.05 in gnomAD were tested for association to age of menopause in women with a premutation using logistic regression, adjusting for repeat and repeat size squared. The quantile-quantile plot indicates there is no population stratification or other oddities of the data (Figure 2.2). As noted in Figure 2.2, no SNP exceeded Bonferroni-adjusted genome-wide significance.

Age at Menopause Polygenic Risk Score Analysis and its association with FXPOI

Using data obtained through a large GWAS study of natural age at menopause (AOM) (Day et al., 2015), we used PRSice (Choi and O'Reilly 2019) to calculate a polygenic risk score (PRS) to test the hypothesis that the polygenic component associated with age at menopause may explain some of the variation in risk for FXPOI. The training set used to derive the PRS for AOM was composed of 69,360 women of European ancestry (Day et al 2015). In that study, 54 SNPs across 44 regions were found to be genome-wide significant, with effect sizes ranging from 0.07 to 0.88 years/allele. Overall, 21% of the variance in age at menopause was explained using 30,000 SNPs with $p<0.05$.

Using PRSice software, we calculated PRS based on SNPs in the discovery dataset at specific p-value threshold sets for association with AOM, adjusting for the first 5 PCs and repeat size and repeat size squared. In this analysis, the maximum variance in risk for FXPOI explained by AOM-associated variants (Nagelkerke's r_2) was 7.5% based on the PRS using SNPs at p-values < 0.002 (Figure 2.3). About 17,000 SNPs have a p-value < 0.002 in the discovery GWAS set, which has 2,407,374 total SNPs. Odds ratios were calculated for quartiles defined by PRS scores of the target dataset.

Only the highest quartile of PRS scores was significant. In the top quartile, the 95% confidence interval spanned a range of 2.12 - 29.35, but the large confidence interval indicates a small sample size in which exact odds ratios are difficult to determine.

Identifying modifying gene candidates with SKAT-O analysis

For the rare variant analysis, we examined variants at MAF< 0.05 and used the kernel based approach SKAT-O that optimizes between burden testing and SKAT models (ref). We adjusted for repeat and repeat size squared. Overall, we interrogated 6,752,810 variants in 25,404 genes. There were no genes that exceeded Bonferroni-adjusted statistical significance that was based on the total number of genes tested. 34 genes passed a threshold of nominal significance at $p < 0.001$ (Table 2.1).

Two additional analyses were conducted on subsets of variants. First, SKAT-O analyses were done filtering on variants located in exon-UTR regions; this included 281,828 variants in 18,975 genes. Second, we filtered on rarer variants at MAF < 0.01; these analyses were based on 4,784,690 variants and 25,346 genes. Sixteen and 31 genes, respectively, passed the nominal statistical significance threshold of $p < 0.001$ (Table 2.2). Genes that passed the threshold from the three analyses were ranked based on p-value, literature evidence of ovarian function or fertility, and having a fly ortholog and TRiP line stock available. Out of the 78 candidate genes from the SKAT-O analyses, 13 genes that met these criteria were chosen for further screening using the *Drosophila* PM model (Table 2.3).

Drosophila fecundity as a whole organism functional study

We first examined whether fecundity was altered in the 90 CGG repeat model compared with controls (Figure 2.4). The controls that were examined included wildtype (OregonR) alone as well as the cross progeny with the *nanos>Gal4* alone and with the 90 CGG repeat, and the cross progeny of the two Bloomington TRiP background lines (Bloomington Stocks # 36303 and # 36304) with the *nanos>Gal4* alone and with the 90 CGG repeat (Figure 2.4). Of the 18 TRiP lines available for knockdown (KD) of the 13 candidate genes, six lines did not produce viable progeny when crossed with the germline-expressing *nanos>Gal4* line (Figure 2.5). Thus, further studies were not performed on these 6 lines, and we proceeded with the remaining 12 lines. Four genotypes per candidate gene were tested: background control with *nanos>Gal4* alone, control with 90 CGG repeat, knockdown of the candidate gene alone, and double mutant containing both 90 CGG repeat and KD of the candidate. A genetic screen with at least three replicate cages, each containing five female flies, for each genotype was performed and the total number of eggs laid was measured over 10 days (see Methods). Out of 9 genes that met the top candidate criteria tested and had viable cross progeny, 3 had the greatest differences between the candidate gene KD alone and candidate gene KD with 90 CGG expressed in the germline (Figure 2.5).

To confirm the apparent differences observed in the initial screen, follow-up experiments of these 3 candidates— *SUMO1*, *KRR1*, and *PDHA2*— were conducted and included at least 10 replicate cages per genotype (Figure 2.6) .Two background

controls for the TRiP lines were used as controls for these replicates (Bloomington #36303 and #36304) depending on which line the candidate gene KD was generated from. Based on the follow-up experiments, we confirmed a significant increase in fecundity for the double mutant containing the 90 CGG repeat and a *SUMO1* KD compared with each of the other genotypes (Figure 2.6). Using a quasipoisson regression model, there was no evidence for an effect of 90 CGG repeat alone or the *SUMO1* KD alone compared with controls; however, the interaction term related to the effect of both mutant genotypes together was statistically significant ($p<0.05$) (Table 2.4). This same pattern was observed for *KRR1*, where the interaction term associated with the effect of the double mutant was statistically significant ($p<0.03$) (Table 2.4). For *PDHA2,* a different pattern was observed. In this case, the effect of the KD itself significantly increased fecundity compared with controls ($p<0.0001$). There was no evidence for an interaction between the *PDHA2* KD and 90 CGG ($p>0.10$) (Table 2.4).

Fecundity of RNA binding proteins

Literature candidates for the RNA sequestration mechanism involved in fragile-X associated disorders (Sellier et al. 2013; 2010) were included in the fecundity screen to compare to the candidates generated in the SKAT-O analysis. Fecundity was measured in the same way as the candidate gene screen. Overall for these RNA binding protein genes, fecundity in both KDs alone and KD/90 CGG double mutants was lower than in their respective background controls. Effects were even more pronounced in some cases. For *CUGB1* RNAi and one of two *Drosha* RNAis, total and near-total loss of

fecundity, respectively, were induced by expression in both KD alone and the corresponding double mutant (Figure 2.7). These results align with the previous studies.

Discussion

In this study, we took the first step to identify genetic variants that play a role in the variable expression of ovarian insufficiency among the women who carry a fragile X PM. Previous work suggested that modifying genetic risk factors do influence age at onset of FXPOI, in addition to the effect of the PM repeat size. Hunter et al. (2008) showed a statistically significant contribution of an additive genetic component to explain risk of FXPOI and Spath et al. (2011) showed an association of the average age of menopause among first degree relatives of women with a PM and the risk for FXPOI, both studies adjusting for the known association of FXPOI with PM repeat size. These findings, combined with studies showing associations of genetic variants for natural age at menopause (Day et al. 2015; Perry et al. 2014; Stolk et al. 2012) and for idiopathic POI (Rossetti et al. 2017; M. Jin, Yu, and Huang 2012) in the general population, motivated us to take a novel strategy that combined WGS and *Drosophila* genetics to identify highly ranked candidate genes that are primed for further study in mammalian systems. We based our studies on women who carried a PM and experienced FXPOI/age at menopause at the extreme tails of the onset distribution: <35 years (cases) and >50 years (controls) of age.

Based on studies that show a significant genetic component related to age at natural menopause, we examined a polygenic risk score (PRS) derived from common

variants associated with lower age at natural menopause identified through a large GWAS study (Day et al. 2015). We found that the PRS explained about 7.5% of the variance in risk for early onset FXPOI, adjusting for PM repeat size and repeat size squared. This result is consistent with our previous findings of an additive genetic component involved in the onset of FXPOI (Hunter et al. 2008; Spath et al. 2011). This suggests that, even on the background of a large single genetic effect, that combined effect of common genetic variants is important as a modifier of severity.

Our next strategy was to examine more rare variants as modifiers of age of onset of FXPOI. We took an untargeted approach and compared WGS variants using several different filtering criteria. 14 genes were highly ranked using gene-set analyses (SKAT-O). Based on a *Drosophila* genetic screen using altered fecundity as an indicator of possible ovarian dysfunction, the germline knockdown of *SUMO1* and *KRR1* were identified as having an interaction with the PM and the germline knockdown of *PDHA2* alone was shown to have an impact on fecundity.

PDHA2 is not directly implicated in ovarian function in its known functions, but variants in *PDHA2* have been linked to male infertility (Sarkar et al. 2019; Yıldırım et al. 2018) and dysregulation of *PDHA2* allows noncanonical expression in somatic tissues (Pinheiro et al. 2016). *KRR1* is an RNA binding protein gene and has previously been identified in studies of polycystic ovarian syndrome (Zheng et al. 2014; Jones and Goodarzi 2016). *SUMO1*'s role in the regulation of granulosa cell apoptosis via sumoylation is particularly interesting because the fragile X-associated PM mouse model was shown to have fewer granulosa cells than wildtype and a faster loss of follicles overall (Hoffman et al. 2012). Reduced ovarian follicular reserve has also been

observed in a mouse model of the premutation (Lu et al. 2012). Phenotypes that have been observed in mouse models of FXPOI include inclusions in granulosa cells and early reduction of the follicular pool (Lu et al. 2012; Conca Dioguardi et al. 2016; Hoffman et al. 2012). Since *SUMO1* is knocked down in this fecundity experiment, it is possible that apoptosis in the follicles of the fly ovaries has been reduced resulting in increased egg laying. However, in this study we aimed to use changes in fecundity in a non-specific reporter, so the mechanism by which reduction in expression of *SUMO1* or *KRR1* combined with the PM could disrupt ovarian function remains a topic for further research.

In addition to taking an untargeted approach, we also tested RNA binding proteins that had previously been associated with the sequestration mechanism associated with fragile X-associated disorders. We found no evidence for these genes playing a modifying role in onset of FXPOI based on our WGS studies. Nonetheless, we further examined the consequences of knocking down each gene using our fecundity screen. The results for the RNA binding protein KDs overall exhibited lower fecundity compared to controls and to the candidate KD with 90 CGG repeat flies. This may be due to a different mechanism of ovarian dysfunction for the candidates found in this study that may not be directly involved in the sequestration method by which the previously established RNA binding proteins confer pathogenesis, but further research would be necessary to fully understand these mechanisms. Ovarian morphology studies in *Drosophila* or in the mouse model may help to better understand differences in mechanism.

Further study of larger cohorts of PM women will generate more candidate genes that help explain the incomplete penetrance and variable expressivity seen in FXPOI. Through fecundity screening as performed in this study, and through potential future molecular analyses, the *Drosophila* PM model serves as a functional assay providing guidance for genes found in WGS analysis. Results from this model serve as a foundation for further research that will be required to determine the mechanism by which these genes interact with the PM.

Tables and Figures

Figure 2.1. Distribution of cohort - A) Distribution of repeat size amongst cases and controls. B) Distribution of age of menopause. All recruited cases experienced menopause before age 35 and range from 16-35. All recruited controls started menopause after age 50.

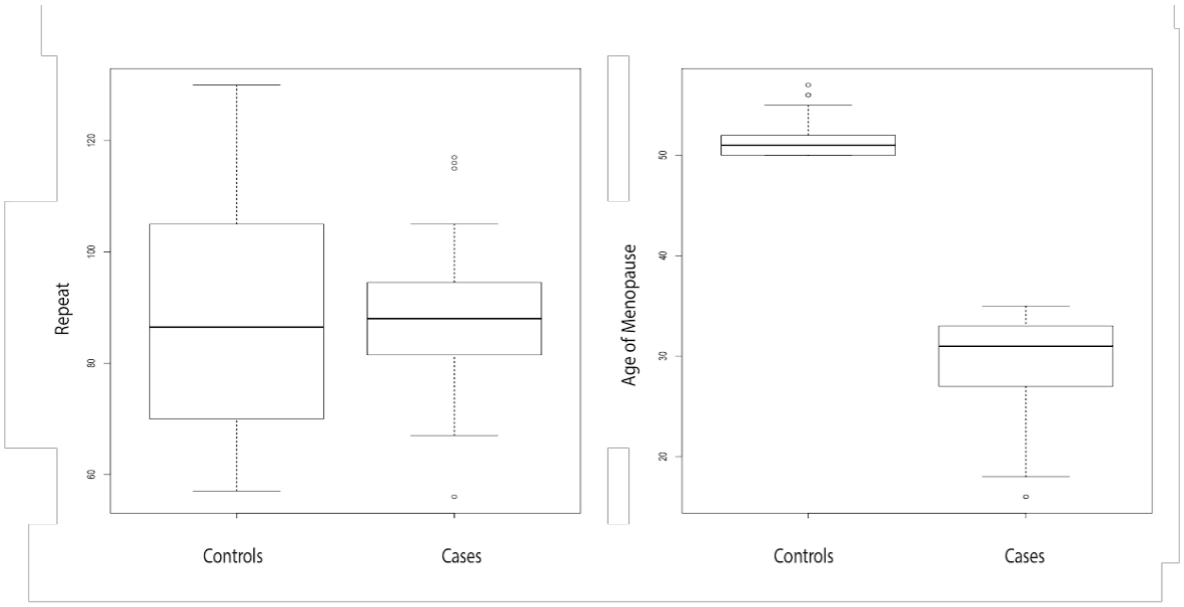


Figure 2.2. Manhattan Plot of SKAT-O results. Each dot represents the p-value for a gene. This is a representative SKAT-O test which is the unified test that weighs between burden test and SKAT for each gene. The red line represents Bonferroni significance, which in this case is less than 1.8x10^-6. The QQ plot shows the results don't deviate from expectation of the null hypothesis. This plot is organized by gene names – it includes 18,000 genes that had rare variants (MAF < 0.01) in cases and controls.

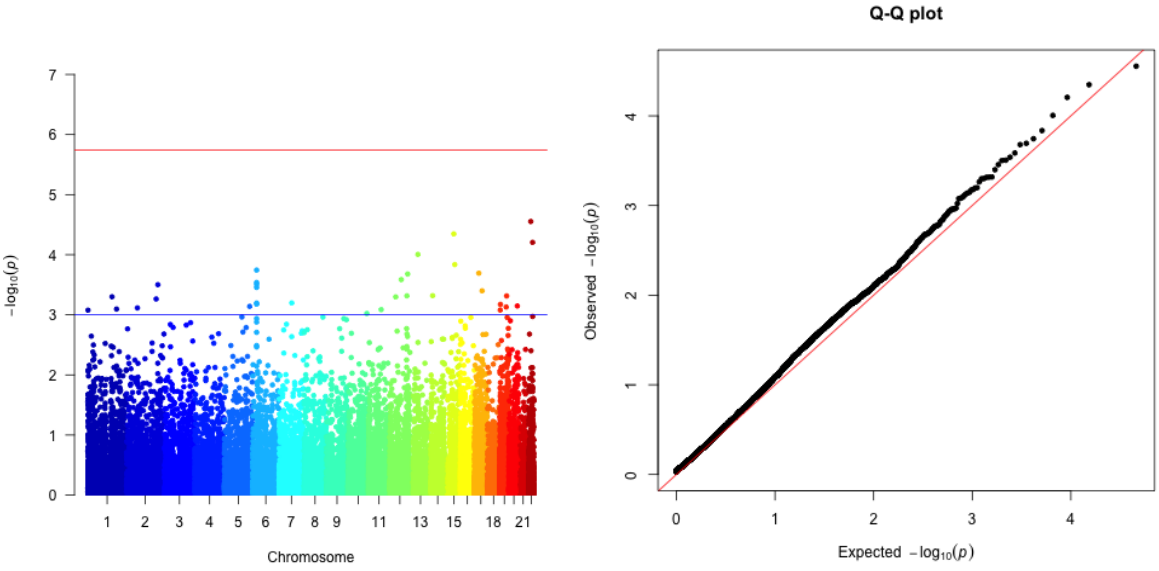


Figure 2.3. PRS analysis reveals a Nagelkerke's R_2 of 7.5% at a threshold of p-values < 0.002 in the discovery set (Day et al. 2015). From left to right, there are increasing numbers of SNPs.

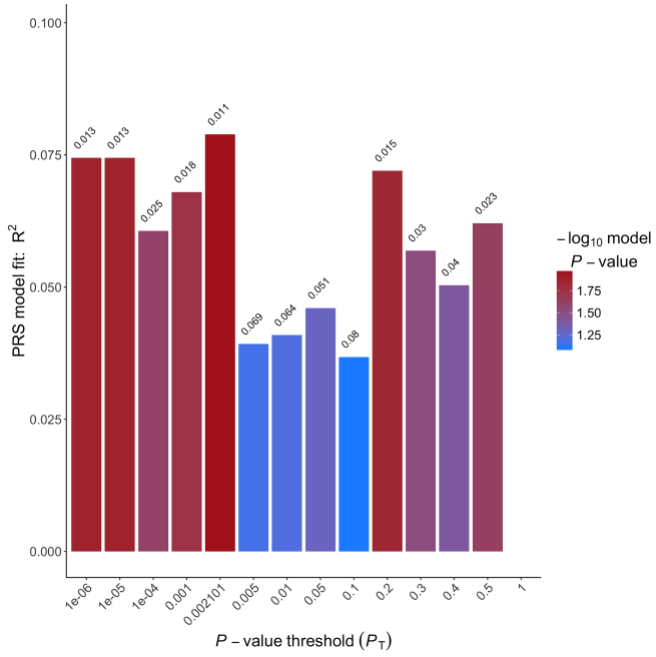


Figure 2.4 Fecundity of Drosophila Controls. Controls included wildtype, wildtype with the cross progeny with *nanos>Gal4* and *nanos>Gal4 90 CGG*, and the cross progeny of the two Bloomington TRiP background lines (Bloomington Stocks #36303 and #36304) with the *nanos>Gal4* alone and with the 90 CGG repeat.

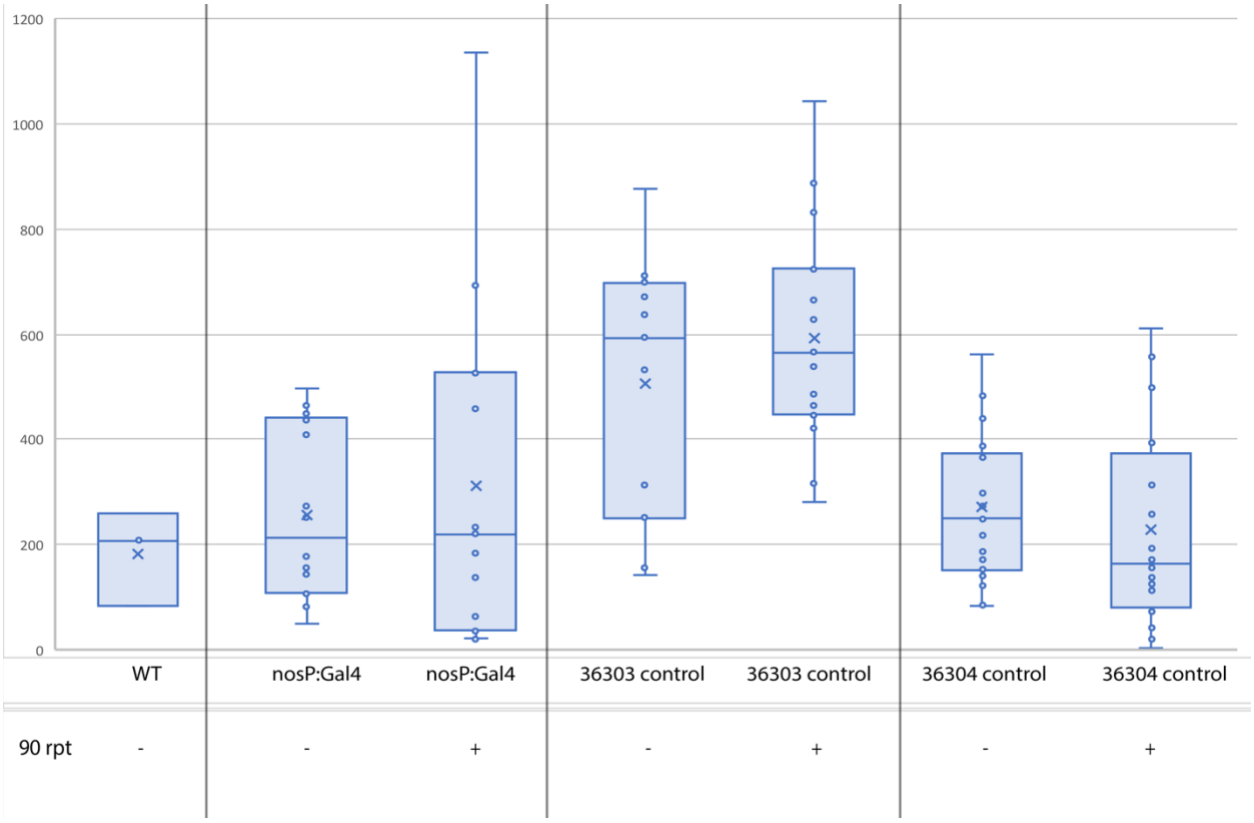


Figure 2.5. Initial screen for top WGS candidate genes. At least three replicates were run for each genotype. Total fecundity represents the egg counts over the 10 days that the experiment ran. The left bar in each pair represents either the control—for the first pair— or KD alone. The right bar in each pair represents either germline expression of 90 CGG alone—for the first pair—or germline expression of 90 CGG along with the candidate gene KD. From this, there were a few possible candidates we compared if the KD alone is different from the double mutant.

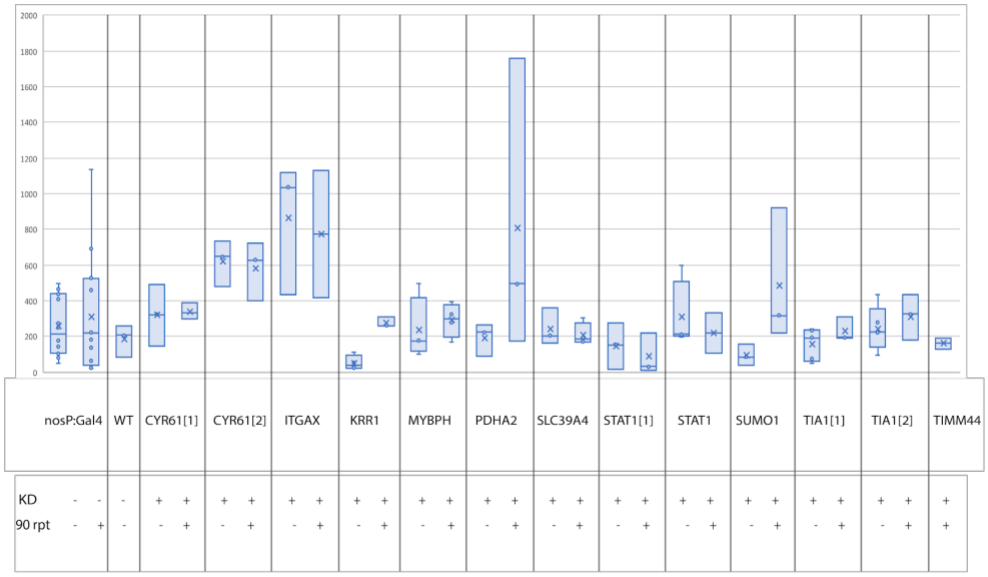


Figure 2.6. Follow-up fecundity testing on top three WGS candidates. Controls included here are the cross progeny of the corresponding Bloomington TRiP background line (Bloomington Stocks # 36303 and # 36304) with the *nanos>Gal4* alone and with the 90 CGG repeat. The number of replicates for each genotype is indicated by 'n.'

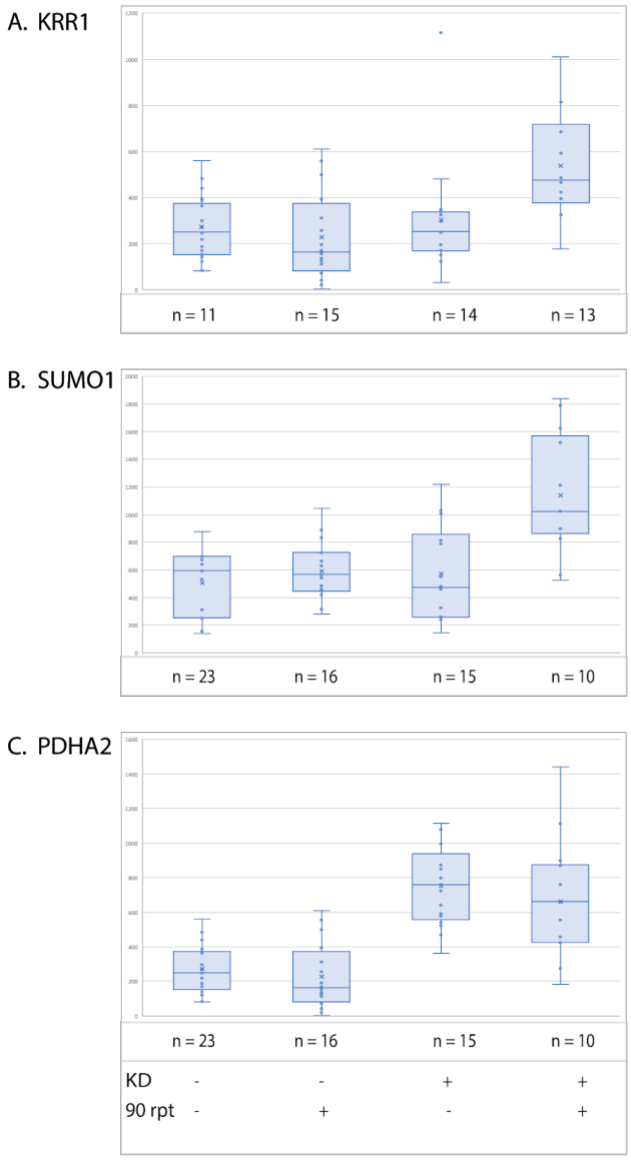


Figure 2.7. Fecundity screen of RNA binding proteins previously associated with Fragile-X associated disorders. Controls included here are the cross progeny of the corresponding Bloomington TRiP background line (Bloomington Stocks #36303 and #36304) with the *nanos>Gal4* alone and with the 90 CGG repeat. There is a trend towards reduction of fecundity overall for these lines.

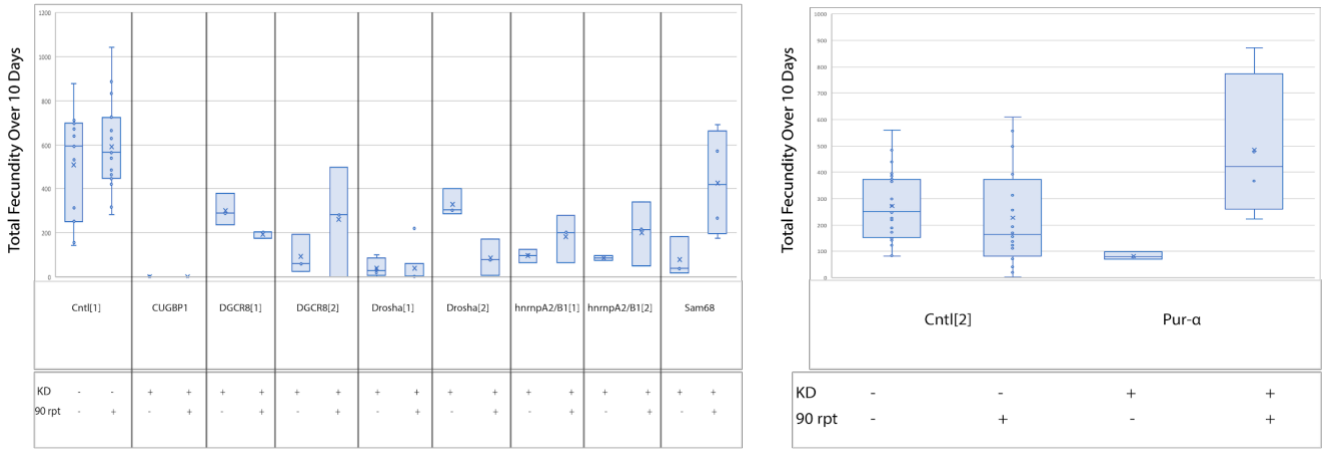


Table 2.1. Bloomington TRiP RNA interference stocks and corresponding human gene orthologs. From a total of 29 ordered stocks, 18 were intended for knockdown of candidate genes identified in this study while the remaining 11 were intended to test previously implicated RNA binding proteins. Of these two groups, 12 stocks and 11 stocks, respectively, produced viable progeny when crossed with *nanos>Gal4* (Bloomington Stock #4442).

Bloomington Stock	**Human Gene**	**Viable cross progeny?**
34806	*DCP2*	no
57029	*KRR1*	yes
31317	*STAT1*	yes
31318	*STAT1*	no
33637	*STAT1*	no
35600	*STAT1*	yes
36125	*SUMO1*	yes
28035	*TIA1*	yes
32472	*TIA1*	yes
31715	*CYR61*	yes
41913	*CYR61*	yes

55281	*SLC39A4*	yes
55345	*PDHA2*	yes
43155	*TIMM44*	no
28535	*ITGAX*	yes
65245	*MYBPH*	yes
60388	*RBPMS2*	no
35759	*TLE6*	no
34896	*Sam68*	yes
36849	*Pur alpha*	yes
67267	*Pur alpha*	yes
31303	*HnrnpA2B1*	yes
32351	*HnrnpA2B1*	yes
35394	*Cugbp1*	yes
44483	*Cugbp1*	yes
27704	*DROSHA*	yes
35233	*DROSHA*	yes
26293	*DGCR8*	yes
33972	*DGCR8*	yes

Table 2.2. Odds ratios for PRS Quartiles. Odds ratios were calculated for different ranges of PRS with 28 or 29 individuals per quartile.

PRS Quartile	Odds Ratio	CI.Upper	CI.Lower	Number of individuals
1	1	1	1	29
2	2.07	6.16	0.69	28

3	1.15	3.57	0.37	28
4	7.89	29.35	2.12	28

Table 2.3. Top candidate genes from SKAT-O analysis. These genes all meet three criteria as they: are top hits in SKAT-O, have fly orthologs, and have roles in ovarian function or fertility.

Gene	P-value SKAT-O	Gene function
CYR61	**3.87E-04**	Involved in angiogenesis processes within reproductive systems (Winterhager and Gellhaus 2014)
DAZL	**4.81E-04**	Germ cell-specific RNA-binding proteins that are implicated in translational regulation of several transcripts (Smorag et al. 2014)
ITGAX	**7.46E-05**	Essential part of the immune cells that are involved in expansion of the cumulus-oocyte complex, release of the ovum from the ovarian follicle, formation of a functional corpus luteum, and enhanced lymphangiogenesis (Cohen-Fredarow et al. 2014)
KRR1	**2.59E-04**	RNA binding protein, associated with PCOS (Zheng et al. 2014)
MYBPH	**4.65E-04**	Reduction of MYBPH mRNA associated with BPA-induced germ cell death (Yin et al. 2016)
PDHA2	**1.27E-04**	Typically only active in male germ cells, but demethylation can lead to expression in other tissues (Pinheiro et al. 2016)
RBPMS2	**5.42E-04**	RNA binding protein that interacts with molecules that are essential to reproduction and egg patterning (Kaufman et al. 2018)
SLC39A4	**6.31E-04**	Transmembrane protein that serves as a zinc uptake protein, zinc levels are heavily associated with fertility (Ford 2004)
STAT1	**5.45E-04**	Regulates granulosa cell function (Benčo et al. 2009)
SUMO1	**3.15E-04**	Involved in apoptosis in granulosa cells (Shao et al. 2006; Liu et al. 2013)
TIA1	**7.68E-04**	mRNA binding protein associated with programmed cell death (Wigington et al. 2015)
TLE6	**8.32E-04**	Maternal effect gene that encodes a member of the subcortical maternal complex in mammalian oocytes, mutations in this gene have associated with sterility in females (Alazami et al. 2015)

Table 2.4. Quasipoisson regression model for top three candidates. The independent variables in the model included presence of the CGG, presence of a knockdown, and the interaction term between those two.

SUMO1

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.2299	0.1591	39.166	<2e-16
CGG	0.1534	0.203	0.756	0.4534
KD	0.1213	0.2071	0.586	0.5608
CGG*KD	0.5314	0.2615	2.032	0.0476

KRR1

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.5987	0.1439	38.912	<2e-16
CGG	-0.1724	0.2368	-0.728	0.4693
KD	0.1185	0.2211	0.536	0.594
CGG*KD	0.7373	0.3291	2.24	0.0288

PDHA2

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.59874	0.13301	42.091	< 2.00E-16
CGG	-0.17242	0.21889	-0.788	0.434

KD	1.01195	0.16246	6.229	3.53E-08
CGG*KD	0.04095	0.2605	0.157	0.876

References

Allen, E. G., A. K. Sullivan, M. Marcus, C. Small, C. Dominguez, M. P. Epstein, K. Charen, W. He, K. C. Taylor, and S. L. Sherman. 2007. “Examination of Reproductive Aging Milestones among Women Who Carry the FMR1 Premutation.” *Human Reproduction (Oxford, England)* 22 (8): 2142–52. https://doi.org/10.1093/humrep/dem148.

Armstrong, Alissa Richmond. 2020. “Drosophila Melanogaster as a Model for Nutrient Regulation of Ovarian Function.” *Reproduction* 159 (2): R69–82. https://doi.org/10.1530/REP-18-0593.

Aumiller, Verena, Almut Graebsch, Elisabeth Kremmer, Dierk Niessing, and Klaus Förstemann. 2012. “Drosophila Pur-α Binds to Trinucleotide-Repeat Containing Cellular RNAs and Translocates to the Early Oocyte.” *RNA Biology* 9 (5): 633–43. https://doi.org/10.4161/rna.19760.

Benčo, A., A. V. Sirotkin, D. Vašíček, S. Pavlová, J. Zemanová, J. Kotwica, K. Darlak, and F. Valenzuela. 2009. “Involvement of the Transcription Factor STAT1 in the Regulation of Porcine Ovarian Granulosa Cell Functions Treated and Not Treated with Ghrelin.” *Reproduction* 138 (3): 553–60. https://doi.org/10.1530/REP-08-0313.

Bione, Silvia, Sara Benedetti, Mara Goegan, Immacolata Menditto, Anna Marozzi,

Maurizio Ferrari, and Daniela Toniolo. 2006. “Skewed X-Chromosome Inactivation Is Not Associated with Premature Ovarian Failure in a Large Cohort of Italian Patients.” *American Journal of Medical Genetics. Part A* 140 (12): 1349–51. https://doi.org/10.1002/ajmg.a.31312.

Choi, Shing Wan, and Paul F. O’Reilly. 2019. “PRSice-2: Polygenic Risk Score Software for Biobank-Scale Data.” *GigaScience* 8 (7). https://doi.org/10.1093/gigascience/giz082.

Cohen-Fredarow, Adva, Ari Tadmor, Tal Raz, Naama Meterani, Yoseph Addadi, Nava Nevo, Inna Solomonov, et al. 2014. “Ovarian Dendritic Cells Act as a Double-Edged Pro-Ovulatory and Anti-Inflammatory Sword.” *Molecular Endocrinology* 28 (7): 1039–54. https://doi.org/10.1210/me.2013-1400.

Conca Dioguardi, Carola, Bahar Uslu, Monique Haynes, Meltem Kurus, Mehmet Gul, De-Qiang Miao, Lucia De Santis, et al. 2016. “Granulosa Cell and Oocyte Mitochondrial Abnormalities in a Mouse Model of Fragile X Primary Ovarian Insufficiency.” *Molecular Human Reproduction* 22 (6): 384–96. https://doi.org/10.1093/molehr/gaw023.

Cooper, G. S., D. P. Sandler, and M. Bohlig. 1999. “Active and Passive Smoking and the Occurrence of Natural Menopause.” *Epidemiology (Cambridge, Mass.)* 10 (6): 771–73.

Daenzer, Jennifer M. I., Rebecca D. Sanders, Darwin Hang, and Judith L. Fridovich-Keil. 2012. “UDP-Galactose 4′-Epimerase Activities toward UDP-Gal and UDP-GalNAc Play Different Roles in the Development of Drosophila Melanogaster.” *PLoS Genetics* 8 (5). https://doi.org/10.1371/journal.pgen.1002721.

Day, Felix R., Katherine S. Ruth, Deborah J. Thompson, Kathryn L. Lunetta, Natalia Pervjakova, Daniel I. Chasman, Lisette Stolk, et al. 2015. “Large-Scale Genomic Analyses Link Reproductive Ageing to Hypothalamic Signaling, Breast Cancer Susceptibility and BRCA1-Mediated DNA Repair.” *Nature Genetics* 47 (11): 1294–1303. https://doi.org/10.1038/ng.3412.

Ennis, Sarah, Daniel Ward, and Anna Murray. 2006. “Nonlinear Association between CGG Repeat Number and Age of Menopause in FMR1 Premutation Carriers.” *European Journal of Human Genetics: EJHG* 14 (2): 253–55. https://doi.org/10.1038/sj.ejhg.5201510.

Ford, Dianne. 2004. “Intestinal and Placental Zinc Transport Pathways.” *Proceedings of the Nutrition Society* 63 (1): 21–29. https://doi.org/10.1079/PNS2003320.

Handa, Vaishali, Tapas Saha, and Karen Usdin. 2003. “The Fragile X Syndrome Repeats Form RNA Hairpins That Do Not Activate the Interferon-inducible Protein Kinase, PKR, but Are Cut by Dicer.” *Nucleic Acids Research* 31 (21): 6243–48. https://doi.org/10.1093/nar/gkg818.

Hoffman, Gloria E., Wei Wei Le, Ali Entezam, Noriyuki Otsuka, Zhi-Bin Tong, Lawrence Nelson, Jodi A. Flaws, John H. McDonald, Sanjeeda Jafar, and Karen Usdin. 2012. “Ovarian Abnormalities in a Mouse Model of Fragile X Primary Ovarian Insufficiency.” *The Journal of Histochemistry and Cytochemistry: Official Journal of the Histochemistry Society* 60 (6): 439–56. https://doi.org/10.1369/0022155412441002.

Hunter, Jessica Ezzell, Michael P. Epstein, Stuart W. Tinker, Krista H. Charen, and Stephanie L. Sherman. 2008. “Fragile X-Associated Primary Ovarian

Insufficiency: Evidence for Additional Genetic Contributions to Severity." *Genetic Epidemiology* 32 (6): 553–59. https://doi.org/10.1002/gepi.20329.

Jin, Peng, Daniela C. Zarnescu, Fuping Zhang, Christopher E. Pearson, John C. Lucchesi, Kevin Moses, and Stephen T. Warren. 2003. "RNA-Mediated Neurodegeneration Caused by the Fragile X Premutation RCGG Repeats in Drosophila." *Neuron* 39 (5): 739–47. https://doi.org/10.1016/s0896-6273(03)00533-6.

Johnston, H. Richard, Pankaj Chopra, Thomas S. Wingo, Viren Patel, International Consortium on Brain and Behavior in 22q11.2 Deletion Syndrome, Michael P. Epstein, Jennifer G. Mulle, Stephen T. Warren, Michael E. Zwick, and David J. Cutler. 2017. "PEMapper and PECaller Provide a Simplified Approach to Whole-Genome Sequencing." *Proceedings of the National Academy of Sciences of the United States of America* 114 (10): E1923–32. https://doi.org/10.1073/pnas.1618065114.

Jones, Michelle R., and Mark O. Goodarzi. 2016. "Genetic Determinants of Polycystic Ovary Syndrome: Progress and Future Directions." *Fertility and Sterility* 106 (1): 25–32. https://doi.org/10.1016/j.fertnstert.2016.04.040.

Kaufman, Odelya H., KathyAnn Lee, Manon Martin, Sophie Rothhämel, and Florence L. Marlow. 2018. "Rbpms2 Functions in Balbiani Body Architecture and Ovary Fate." *PLOS Genetics* 14 (7): e1007489. https://doi.org/10.1371/journal.pgen.1007489.

Kenneson, A., F. Zhang, C. H. Hagedorn, and S. T. Warren. 2001. "Reduced FMRP and Increased FMR1 Transcription Is Proportionally Associated with CGG Repeat

Number in Intermediate-Length and Premutation Carriers." *Human Molecular Genetics* 10 (14): 1449–54. https://doi.org/10.1093/hmg/10.14.1449.

Liu, Xiao-Ming, Fei-Fei Yang, Yi-Feng Yuan, Rui Zhai, and Li-Jun Huo. 2013. "SUMOylation of Mouse P53b by SUMO-1 Promotes Its pro-Apoptotic Function in Ovarian Granulosa Cells." *PloS One* 8 (5): e63680. https://doi.org/10.1371/journal.pone.0063680.

Lu, Cuiling, Li Lin, Huiping Tan, Hao Wu, Stephanie L. Sherman, Fei Gao, Peng Jin, and Dahua Chen. 2012. "Fragile X Premutation RNA Is Sufficient to Cause Primary Ovarian Insufficiency in Mice." *Human Molecular Genetics* 21 (23): 5039–47. https://doi.org/10.1093/hmg/dds348.

Nelson, Lawrence M. 2009. "Clinical Practice. Primary Ovarian Insufficiency." *The New England Journal of Medicine* 360 (6): 606–14. https://doi.org/10.1056/NEJMcp0808697.

Perry, John R. B., Yi-Hsiang Hsu, Daniel I. Chasman, Andrew D. Johnson, Cathy Elks, Eva Albrecht, Irene L. Andrulis, et al. 2014. "DNA Mismatch Repair Gene MSH6 Implicated in Determining Age at Natural Menopause." *Human Molecular Genetics* 23 (9): 2490–97. https://doi.org/10.1093/hmg/ddt620.

Pieretti, M., F. P. Zhang, Y. H. Fu, S. T. Warren, B. A. Oostra, C. T. Caskey, and D. L. Nelson. 1991. "Absence of Expression of the FMR-1 Gene in Fragile X Syndrome." *Cell* 66 (4): 817–22. https://doi.org/10.1016/0092-8674(91)90125-i.

Pinheiro, Ana, Maria João Silva, Hana Pavlu-Pereira, Cristina Florindo, Madalena Barroso, Bárbara Marques, Hildeberto Correia, et al. 2016. "Complex Genetic Findings in a Female Patient with Pyruvate Dehydrogenase Complex Deficiency:

Null Mutations in the PDHX Gene Associated with Unusual Expression of the Testis-Specific PDHA2 Gene in Her Somatic Cells." *Gene* 591 (2): 417–24. https://doi.org/10.1016/j.gene.2016.06.041.

Primerano, Beatrice, Flora Tassone, Randi J. Hagerman, Paul Hagerman, Francesco Amaldi, and Claudia Bagni. 2002. "Reduced FMR1 MRNA Translation Efficiency in Fragile X Patients with Premutations." *RNA (New York, N.Y.)* 8 (12): 1482–88.

R: A Language and Environment for Statistical Computing. 2017. Vienna, Austria: R Foundation for Statistical Computing.

Rodriguez-Revenga, Laia, Irene Madrigal, Celia Badenas, Mar Xunclà, Loli Jiménez, and Montserrat Milà. 2009. "Premature Ovarian Failure and Fragile X Female Premutation Carriers: No Evidence for a Skewed X-Chromosome Inactivation Pattern." *Menopause (New York, N.Y.)* 16 (5): 944–49. https://doi.org/10.1097/gme.0b013e3181a06a37.

Sarkar, Saumya, Kumar Mohanty Sujit, Vertika Singh, Rajesh Pandey, Sameer Trivedi, Kiran Singh, Gopal Gupta, and Singh Rajender. 2019. "Array-Based DNA Methylation Profiling Reveals Peripheral Blood Differential Methylation in Male Infertility." *Fertility and Sterility* 112 (1): 61-72.e1. https://doi.org/10.1016/j.fertnstert.2019.03.020.

Schindelin, Johannes, Curtis T. Rueden, Mark C. Hiner, and Kevin W. Eliceiri. 2015. "The ImageJ Ecosystem: An Open Platform for Biomedical Image Analysis." *Molecular Reproduction and Development* 82 (7–8): 518–29. https://doi.org/10.1002/mrd.22489.

Sellier, Chantal, Ronald A. M. Buijsen, Fang He, Sam Natla, Laura Jung, Philippe

Tropel, Angeline Gaucherot, et al. 2017. “Translation of Expanded CGG Repeats into FMRpolyG Is Pathogenic and May Contribute to Fragile X Tremor Ataxia Syndrome.” *Neuron* 93 (2): 331–47. https://doi.org/10.1016/j.neuron.2016.12.016.

Sellier, Chantal, Fernande Freyermuth, Ricardos Tabet, Tuan Tran, Fang He, Frank Ruffenach, Violaine Alunni, et al. 2013. “Sequestration of DROSHA and DGCR8 by Expanded CGG RNA Repeats Alters MicroRNA Processing in Fragile X-Associated Tremor/Ataxia Syndrome.” *Cell Reports* 3 (3): 869–80. https://doi.org/10.1016/j.celrep.2013.02.004.

Sellier, Chantal, Frédérique Rau, Yilei Liu, Flora Tassone, Renate K. Hukema, Renata Gattoni, Anne Schneider, et al. 2010. “Sam68 Sequestration and Partial Loss of Function Are Associated with Splicing Alterations in FXTAS Patients.” *The EMBO Journal* 29 (7): 1248–61. https://doi.org/10.1038/emboj.2010.21.

Shao, Ruijin, Emilia Rung, Birgitta Weijdegård, and Håkan Billig. 2006. “Induction of Apoptosis Increases SUMO-1 Protein Expression and Conjugation in Mouse Periovulatory Granulosa Cells in Vitro.” *Molecular Reproduction and Development* 73 (1): 50–60. https://doi.org/10.1002/mrd.20386.

Sherman, S. L. 2000. “Premature Ovarian Failure in the Fragile X Syndrome.” *American Journal of Medical Genetics* 97 (3): 189–94. https://doi.org/10.1002/1096-8628(200023)97:3<189::AID-AJMG1036>3.0.CO;2-J.

Sherman, Stephanie L, Eliza C Curnow, Charles A Easley, Peng Jin, Renate K Hukema, Maria Isabel Tejada, Rob Willemsen, and Karen Usdin. 2014. “Use of Model Systems to Understand the Etiology of Fragile X-Associated Primary

Ovarian Insufficiency (FXPOI)." *Journal of Neurodevelopmental Disorders* 6 (1): 26. https://doi.org/10.1186/1866-1955-6-26.

Smorag, L, X Xu, W Engel, and Pantakani Pantakani. 2014. "The Roles of DAZL in RNA Biology and Development." Wiley Interdisciplinary Reviews. RNA. Wiley Interdiscip Rev RNA. August 2014. https://doi.org/10.1002/wrna.1228.

Sofola, Oyinkan A., Peng Jin, Yunlong Qin, Ranhui Duan, Huijie Liu, Maria de Haro, David L. Nelson, and Juan Botas. 2007. "RNA-Binding Proteins HnRNP A2/B1 and CUGBP1 Suppress Fragile X CGG Premutation Repeat-Induced Neurodegeneration in a Drosophila Model of FXTAS." *Neuron* 55 (4): 565–71. https://doi.org/10.1016/j.neuron.2007.07.021.

Spath, Marian A., Ton B. Feuth, Arie P.T. Smits, Helger G. Yntema, Didi D.M. Braat, Chris M.G. Thomas, Ad Geurts van Kessel, Stephanie L. Sherman, and Emily G. Allen. 2011. "Predictors and Risk Model Development for Menopausal Age in Fragile x Premutation Carriers." *Genetics in Medicine : Official Journal of the American College of Medical Genetics* 13 (7): 643–50. https://doi.org/10.1097/GIM.0b013e31821705e5.

Spath, Marian A., Willy N. Nillesen, Arie P. T. Smits, Ton B. Feuth, Didi D. M. Braat, Ad Geurts van Kessel, and Helger G. Yntema. 2010. "X Chromosome Inactivation Does Not Define the Development of Premature Ovarian Failure in Fragile X Premutation Carriers." *American Journal of Medical Genetics. Part A* 152A (2): 387–93. https://doi.org/10.1002/ajmg.a.33243.

Stolk, Lisette, John RB Perry, Daniel I Chasman, Chunyan He, Massimo Mangino, Patrick Sulem, Maja Barbalic, et al. 2012. "Meta-Analyses Identify 13 Novel Loci

Associated with Age at Menopause and Highlights DNA Repair and Immune Pathways." *Nature Genetics* 44 (3): 260–68. https://doi.org/10.1038/ng.1051.

Sullivan, A. K., M. Marcus, M. P. Epstein, E. G. Allen, A. E. Anido, J. J. Paquin, M. Yadav-Shah, and S. L. Sherman. 2005. "Association of FMR1 Repeat Size with Ovarian Dysfunction." *Human Reproduction (Oxford, England)* 20 (2): 402–12. https://doi.org/10.1093/humrep/deh635.

Tassone, F., and P. J. Hagerman. 2003. "Expression of the FMR1 Gene." *Cytogenetic and Genome Research* 100 (1–4): 124–28. https://doi.org/10.1159/000072846.

Tassone, F., R. J. Hagerman, A. K. Taylor, L. W. Gane, T. E. Godfrey, and P. J. Hagerman. 2000. "Elevated Levels of FMR1 MRNA in Carrier Males: A New Mechanism of Involvement in the Fragile-X Syndrome." *American Journal of Human Genetics* 66 (1): 6–15. https://doi.org/10.1086/302720.

Tassone, F., R. J. Hagerman, A. K. Taylor, J. B. Mills, S. W. Harris, L. W. Gane, and P. J. Hagerman. 2000. "Clinical Involvement and Protein Expression in Individuals with the FMR1 Premutation." *American Journal of Medical Genetics* 91 (2): 144–52. https://doi.org/10.1002/(sici)1096-8628(20000313)91:2<144::aid-ajmg14>3.0.co;2-v.

Tejada, Maria-Isabel, Eva García-Alegría, Amaia Bilbao, Cristina Martínez-Bouzas, Elena Beristain, Marisa Poch, Maria A. Ramos-Arroyo, et al. 2008. "Analysis of the Molecular Parameters That Could Predict the Risk of Manifesting Premature Ovarian Failure in Female Premutation Carriers of Fragile X Syndrome." *Menopause (New York, N.Y.)* 15 (5): 945–49. https://doi.org/10.1097/gme.0b013e3181647762.

Todd, Peter K., Seok Yoon Oh, Amy Krans, Fang He, Chantal Sellier, Michelle Frazer, Abigail J. Renoux, et al. 2013. “CGG Repeat-Associated Translation Mediates Neurodegeneration in Fragile X Tremor Ataxia Syndrome.” *Neuron* 78 (3): 440–55. https://doi.org/10.1016/j.neuron.2013.03.026.

Vabre, Pauline, Nicolas Gatimel, Jessika Moreau, Véronique Gayrard, Nicole Picard-Hagen, Jean Parinaud, and Roger D. Leandri. 2017. “Environmental Pollutants, a Possible Etiology for Premature Ovarian Insufficiency: A Narrative Review of Animal and Human Data.” *Environmental Health* 16 (April). https://doi.org/10.1186/s12940-017-0242-4.

Welt, C. K., P. C. Smith, and A. E. Taylor. 2004. “Evidence of Early Ovarian Aging in Fragile X Premutation Carriers.” *The Journal of Clinical Endocrinology and Metabolism* 89 (9): 4569–74. https://doi.org/10.1210/jc.2004-0347.

Wigington, Callie P., Jeenah Jung, Emily A. Rye, Sara L. Belauret, Akahne M. Philpot, Yue Feng, Philip J. Santangelo, and Anita H. Corbett. 2015. “Post-Transcriptional Regulation of Programmed Cell Death 4 (PDCD4) MRNA by the RNA-Binding Proteins Human Antigen R (HuR) and T-Cell Intracellular Antigen 1 (TIA1).” *The Journal of Biological Chemistry* 290 (6): 3468–87. https://doi.org/10.1074/jbc.M114.631937.

Winterhager, Elke, and Alexandra Gellhaus. 2014. “The Role of the CCN Family of Proteins in Female Reproduction.” *Cellular and Molecular Life Sciences* 71 (12): 2299–2311. https://doi.org/10.1007/s00018-014-1556-9.

Yang, Lele, Ranhui Duan, Dongsheng Chen, Jun Wang, Dahua Chen, and Peng Jin. 2007. “Fragile X Mental Retardation Protein Modulates the Fate of Germline

Stem Cells in Drosophila." *Human Molecular Genetics* 16 (15): 1814–20. https://doi.org/10.1093/hmg/ddm129.

Yang, Yingyue, Shunliang Xu, Laixin Xia, Jun Wang, Shengmei Wen, Peng Jin, and Dahua Chen. 2009. "The Bantam MicroRNA Is Associated with Drosophila Fragile X Mental Retardation Protein and Regulates the Fate of Germline Stem Cells." *PLoS Genetics* 5 (4): e1000444. https://doi.org/10.1371/journal.pgen.1000444.

Yıldırım, Yeşerin, Toufik Ouriachi, Ute Woehlbier, Wahiba Ouahioune, Mahmut Balkan, Sajid Malik, and Aslıhan Tolun. 2018. "Linked Homozygous BMPR1B and PDHA2 Variants in a Consanguineous Family with Complex Digit Malformation and Male Infertility." *European Journal of Human Genetics: EJHG* 26 (6): 876–85. https://doi.org/10.1038/s41431-018-0121-7.

Zheng, Sanduo, Pengfei Lan, Ximing Liu, and Keqiong Ye. 2014. "Interaction between Ribosome Assembly Factors Krr1 and Faf1 Is Essential for Formation of Small Ribosomal Subunit in Yeast." *The Journal of Biological Chemistry* 289 (33): 22692–703. https://doi.org/10.1074/jbc.M114.584490.

III. Identifying genetic factors that contribute to the increased risk of congenital heart defects in infants with Down syndrome

Cristina Trevino[1], Aaron M. Holleman[1], Holly Corbitt, Cheryl L. Maslen, Tracie C. Rosser, David J. Cutler, H. Richard Johnston, Benjamin L. Rambo-Martin, Jai Oberoi, Kenneth J. Dooley, George Capone, Roger H. Reeves, Heather J. Cordell, Bernard D. Keavney, A.J. Agopian, Elizabeth Goldmuntz, Peter Gruber, Clifford L. Cua, Jennifer G. Mulle, Michael E. Zwick, Michael P. Epstein, Stephanie L. Sherman
[1]Co-first authors

ABSTRACT

Atrioventricular septal defects (AVSD) are a severe congenital heart defect present in individuals with Down syndrome (DS) at a >2,000-fold increased prevalence compared to the general population. This study aimed to identify risk-associated genes and pathways and to examine a potential polygenic contribution to AVSD in DS. We analyzed a total cohort of 702 individuals with DS and with or without AVSD, with genomic data from whole exome sequencing, whole genome sequencing, and/or array-based imputation. We utilized sequence kernel association testing and polygenic risk score (PRS) methods to examine rare and common variants. Our findings suggest that the Notch pathway, particularly *NOTCH4*, as well as genes involved in the ciliome including *CEP290* may play a role in AVSD in DS. These pathways have also been implicated in DS-associated AVSD in prior studies. A polygenic component for AVSD in DS has not been examined previously. Using the largest discovery GWAS of congenital heart defects available (2,594 cases and 5,159 controls; all general population samples), we found PRS to be associated with AVSD with odds ratios ranging from 1.2 to 1.3 per standard deviation increase in PRS and corresponding liability r_2 values of approximately 1%, suggesting at least a small polygenic contribution to DS-associated

AVSD. Future studies with larger sample sizes will improve identification and quantification of genetic contributions to AVSD in DS.

INTRODUCTION

Congenital heart defects (CHD) are present in over 40% of infants with Down syndrome (DS), with the vast majority being septal defects (Ferencz et al. 1989). Among septal defects, atrioventricular septal defects (AVSD) are the most severe, requiring surgery early in life. Approximately 20% of those with DS have an AVSD, compared to only 1 in 10,000 in the non-DS population (Hartman et al. 2011). This >2,000-fold increase in AVSD prevalence strongly suggests that trisomy 21 and resulting dysregulation of the genome significantly increase the risk for this disorder. Furthermore, it is likely that other genetic variation across the genome contributes to DS-associated AVSD; however, identification remains elusive. Efforts to clarify the genetic basis of AVSD in DS are important, as improved understanding of genetic causes may inform future work that facilitates a decrease in AVSD burden among the DS community; moreover, it has potential to shed light on fundamental biology relevant to the formation of CHD generally, which could yield benefits that extend beyond those with DS.

There have been several studies of the role of common variants in DS-associated AVSD, including the largest GWAS to date with 210 complete AVSD cases with DS (DS+AVSD; diagnosed with full trisomy 21) and 242 controls with DS and structurally normal hearts (DS+NH). These studies have not identified any common variants (SNPs or CNVs) exceeding genome-wide significance, despite adequate

sample sizes for detecting common variants with large effect sizes (Ramachandran, Zeng, et al. 2015; Ramachandran, Mulle, et al. 2015; Rambo-Martin et al. 2018). This suggests that large-effect common variants (e.g., odds ratios > 2.0) do not play a significant role in DS-associated AVSD. However, low to moderate-effect common variants including SNVs and structural variants may be contributing to risk, perhaps in a cumulative way (Sailani et al. 2013).

Rare variant studies of AVSD have yielded some positive results, both among those with DS and those with non-syndromic AVSD. In a targeted sequencing study of 26 AVSD candidate genes among 141 DS+AVSD cases and 141 DS+NH controls, rare variants with predicted deleterious effects were found to be enriched in cases for genes involved in the vascular endothelial growth factor pathway (Ackerman et al. 2012). A CNV analysis of the aforementioned 210 DS+AVSD cases and 242 DS+NH controls identified a suggestive enrichment of large rare deletions in ciliome genes among cases (Ramachandran, Mulle, et al. 2015). A more recent study of 198 DS+AVSD cases and 211 DS+NH controls, a subset of those analyzed by Ramachandran et al., investigated CNVs on the trisomic chromosome 21. The investigators found controls self-identifying as African-American to have more bases covered by rare deletions than African-American cases, while cases self-identifying as Caucasian had more genes intersected by rare duplications than Caucasian controls (Rambo-Martin et al. 2018).

Among those with nonsyndromic AVSD, a rare variant exome study (MAF < 0.01) involving 13 parent-offspring trios of probands and 112 unrelated controls revealed cases to be enriched for missense variants in *NR2F2,* a gene that encodes for a nuclear receptor that is part of a steroid hormone superfamily and has been shown to play a role

in heart development in mouse studies (Al Turki et al. 2014; Lin et al. 2012; Pereira et al. 1999). These findings suggest that rare variants may play an important role in DS-associated AVSD, warranting further study with larger sample sizes.

The limited success in identifying AVSD-associated sequence variants, both common and rare, may be due to small sample sizes that inhibit discovery. It is also possible that the genetic architecture of CHD is more complex than originally hypothesized. For common complex disorders such as schizophrenia and cardiovascular disease, it is now understood that there is a polygenic component to risk, whereby hundreds or thousands of common variants each incrementally increase risk for the phenotype (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014; Khera et al. 2018). AVSD may similarly have a polygenic component contributing to risk, which could be especially relevant when combined with a genomic background in which many genes are dysregulated due to trisomy 21. This polygenic component can be quantified using a polygenic risk score (PRS) methodology, which examines the extent to which common variants (MAF > 0.05) may be collectively contributing to a phenotype.

Rare variants may also play an important role in AVSD. When working with a rare disorder such as DS+AVSD, it is essential to maximize the power of small sample sizes, as it is anticipated that many mutations will be private or ultra-rare. This notion supports the use of burden tests or the sequence kernel association test (SKAT) (Wu et al. 2010), which group rare variants into those occurring in genes or pathways. Use of the optimal unified test (SKAT-O) maximizes the advantages of both types of combined variant testing (Lee et al. 2012) by modeling both SKAT and the burden test for each

defined variant set and finding the optimal linear combination of both tests, thus optimizing power for the test. SKAT-O can be employed for analysis of common variants within genes and pathways as well as for rare variant testing (Ionita-Laza et al. 2013).

While the PRS and SKAT-O approaches are not designed to pinpoint individual genetic variants as associated with the target phenotype, they provide insight into the genetic underpinnings of the trait of interest that complements standard analyses of individual genetic variation. With this in mind, we implemented the PRS approach and SKAT-O in order to optimally examine whether common and rare variants may be contributing to DS-associated AVSD and to identify those genes and pathways that may be most relevant to this phenotype.

METHODS

Subjects

Participant samples were obtained through two methods: the Down Syndrome Heart (Project) (DSHP) and the Pediatric Cardiac Genomic Consortium (PCGC). Through the DSHP, probands with trisomy 21 with and without heart defects were identified at multiple sites in the U.S., as described previously (Ramachandran, Mulle, et al. 2015; Ramachandran, Zeng, et al. 2015). PCGC probands were collected from multiple sites in the U.S. and U.K (Hoang et al. 2018; “The Congenital Heart Disease Genetic Network Study” 2013a). The PCGC parent study recruited probands with heart defects, and in these analyses we used data for the subset of those with heart defects who also had trisomy 21. Thus, the inclusion criteria were similar for all sites for both cohorts. All

participants were diagnosed with full or translocation trisomy 21, with the vast majority documented by karyotype or medical records. Cases in both cohorts were defined as individuals with trisomy 21 and a complete, balanced AVSD diagnosed by echocardiogram or surgical reports. For the DSHP, a single cardiologist (K. Dooley) defined cases based on medical records. For the PCGC, details of CHD review are described in Hoang et al. 2018. Controls were classified as individuals from the DSHP with trisomy 21 and a structurally normal heart, patent foramen ovale, or patent ductus arteriosus (DS+NH). The majority of controls were defined based on echocardiograms. The PCGC case samples (n=47) were only included in the WGS dataset; none overlap with those obtained through the DSHP. Table 3.1 provides the sample sizes of the WES and WGS datasets.

Institutional review boards at each enrolling institution approved protocols and informed consent was obtained from a custodial parent for each participant.

Whole exome sequencing

Whole exome sequencing (WES) was performed on 190 DS+AVSD cases and 138 DS+NH controls by the NHLBI Resequencing and Genotyping Service at the University of Washington. FASTQ files from single-ended whole exome sequencing (WES) were mapped and variants were called with Emory's PEMapper and PECaller, respectively (Johnston et al. 2017). Variants were annotated using Bystro (Kotlar et al. 2018) (http://bystro.io). A total of 331,935 single nucleotide variants (SNVs) were detected by whole exome sequencing across the 190 cases and 138 controls. Mean

coverage depth ± standard deviation (sd) of exome sequencing was 1.57 ± 0.39 for samples and mean transition/transversion ratio ± sd was 2.7 ± 0.15.

Sample failures were addressed by removing any individuals missing > 1% genotypes or failing PLINK1.9's sex check (based on F statistics for X chromosome heterozygosity, which were also used to impute sex on individuals missing sex data) (Weir and Cockerham 1984; Chang et al. 2015, Purcell and Chang). These filters identified no samples for exclusion. Variant filters included removing SNVs with missingness > 10% and those failing the exact test for HWE at a p-value < 10^{-6}.

We then performed principal component analysis (PCA), using PLINK1.9. We used common SNPs (MAF > 0.05) and pruned SNPs in linkage disequilibrium with an r^2 > 0.2, stepping along five SNPs at a time within 50kb windows. Through three rounds of PCA we identified a total of 28 outlier samples for removal. Following QC, the WES dataset contained 300 samples (174 cases, 126 controls) and 330,287 SNVs for analysis.

Whole genome sequencing

Paired-ended whole genome sequencing (WGS) was performed on 169 DS+AVSD cases and 39 DS+NH controls by Hudson Alpha (Huntsville, AL) to a target depth of 30x. Raw FASTQ data were mapped and variants were called using PEMapper and PECaller, respectively, and variants were annotated using Bystro. In total, 12,302,231 SNVs were detected by whole genome sequencing across 169 cases and 39 controls. Mean coverage depth ± sd was 30.2 ± 4.1. Mean transition/transversion ratio ± sd was 2.05 ± 0.007.

Sample failures were addressed by removing samples with theta < 3 sd below mean theta, transition/transversion ratio < 3 sd below mean transition/transversion ratio, heterozygosity/homozygosity ratio > 4 sd above mean heterozygosity/homozygosity ratio, missing > 1% of genotypes, or failing PLINK1.9's sex check, and excluding one sample from each pair of related samples (based on PLINK1.9 PI_HAT > 0.1875). These steps resulted in the removal of 16 poor quality WGS samples. We also excluded one WGS sample identified as a duplicate of a WES sample. Variant QC involved removing SNVs with missingness > 10%, and those failing the exact test for HWE among cases and controls combined at a p-value < 10^{-12}.

We performed PCA in the same manner as described for the WES dataset. Three rounds of PCA identified a total of 16 additional WGS samples as outliers for removal. After these QC steps, the WGS dataset contained 175 samples (148 cases and 27 controls) and 12,279,101 variants.

Samples with imputed genotypes based on microarray

Affymetrix Genome-Wide Human SNP 6.0 array genotype data were available for 459 DS samples (211 DS+AVSD cases, 248 DS+NH controls), including 198 (100 cases, 98 controls) of the 328 WES samples and 95 (all cases) of the 208 WGS samples described above. Array data for these 459 DS samples were originally generated and analyzed in the prior GWAS and CNV analysis of DS-associated AVSD (Ramachandran, Mulle, et al. 2015; Ramachandran, Zeng, et al. 2015). We applied standard GWAS QC and PCA procedures to these data (see Supplemental Methods for details) using PLINK1.9 and R (version 3.4.1) (*R: A Language and Environment for Statistical Computing* 2017), which yielded a dataset with 207 cases and 234 controls,

all of European ancestry, and 612,125 autosomal SNPs (excluding the trisomic chromosome 21).

For these samples, we performed genotype imputation using the Michigan Imputation Server (Das et al. 2016). Genotype imputation was based on the Haplotype Reference Consortium (HRC) panel (version r1-1 2016) (McCarthy et al. 2016), which includes 32,470 samples predominantly of European ancestry. The post-imputation files included 38,596,402 autosomal variants (all SNPs). Mean correlation between true and imputed genotypes for the ~600,000 genotyped SNPs was 0.990, suggesting high quality imputation. For this dataset (referred to as the "imputed dataset"), we excluded variants with MAF < 0.01, those missing for more than 2% of samples, those with a maximum imputed genotype probability < 0.80, and those with imputation r_2 < 0.80.

We then applied standard GWAS QC to the imputed dataset (see Supplemental Methods for details). We also removed variants with A/T, T/A, C/G, and G/C alleles which can be difficult to match between datasets due to strand ambiguity; this was done in preparation for merging this imputed dataset with unique WGS samples, to create a larger sample for the PRS analyses (these steps are explained in detail in the PRS Analysis section). This left an imputed dataset with 440 samples (206 cases, 234 controls) and 5,079,537 autosomal SNPs.

Analyses

SKAT-O variant analysis

All variants in both the WES and WGS datasets were filtered using Bystro (Kotlar et al. 2018) to include only exonic and UTR regions of the genome for any transcript as defined by RefSeq (hg38). Variants in multiple overlapping genes were included in each gene separately. They were further filtered using Bystro to only include SNVs. Separate analyses were conducted for variants within the specified minor allele frequency (MAF) categories (Table 3.2). GnomAD MAFs were used to define eligible variants for rare or common analyses. In order to capture variants at a very low population MAF in our dataset, variants missing from gnomAD were included in the rare analysis. As an additional step to ensure those missing variants were truly at a low MAF, the variants that were missing allele frequency information in gnomAD were filtered based on the WES or WGS dataset at MAF < 0.02. In contrast, for the ultra-rare analysis we excluded the variants missing from gnomAD in order to test whether well-defined ultra-rare variants could be driving the top rare results. MAFs from gnomAD were used as weights in the rare and ultra-rare variant analyses (Table 3.2). Records where the reference allele was the minor allele were excluded. Variants in chromosome 21 were excluded. SKAT-O testing was done using the SKAT package in R using sex and the first five principal components of ancestry as covariates. For the SKAT-O analysis, the WES dataset was analyzed first. Genes with a resulting $p < 0.001$ (small-sample adjusted) were then evaluated in the WGS dataset by SKAT-O. The genes with the lowest p-value were evaluated as candidate genes. Candidate genes were checked for cardiac phenotypes and heart expression using GTEx.

Rare variant pathway analysis

As a follow-up to the gene-based SKAT-O tests, two pathways that our top candidate genes belong to were evaluated due to their reported involvement in heart development: 1) ciliome gene set (van Dam et al. 2013) and 2) Notch pathway (Kanehisa 2000). Each was evaluated as a single gene set. For the ciliome, 3,573 SNVs identified in 301 genes found in the van Dam Ciliome gene list (van Dam et al. 2013) were evaluated. For the Notch pathway, 487 SNVs identified in 48 genes in the Notch pathway defined by KEGG (Kanehisa 2000) were analyzed. All variants within a defined MAF-filtered category (common, rare, and ultra-rare) were analyzed with SKAT-O.

Polygenic Risk Score (PRS) analyses

We have grouped the PRS analyses into primary and secondary analyses. The primary analyses had the goal of examining the genome-wide polygenic contribution to DS-associated AVSD, while the secondary analyses had the goal of estimating the additional polygenic contribution specifically due to the trisomic chromosome 21. These primary and secondary PRS analyses utilized slightly different target datasets and slightly different processes for generating and analyzing the PRS (as described below), but employed the same discovery datasets for weighting alleles in the PRS.

Target dataset for primary analyses

The target sample for our primary PRS analyses included 245 DS+AVSD cases and 242 DS+NH controls and represents a combination of the WGS and imputed datasets. We prepared the WGS dataset (175 samples) for merger with the imputed dataset (440 samples) by removing variants with MAF < 0.01, those missing for > 2% of samples, and indels (filters which had already been applied to the imputed dataset). We then merged these datasets and subjected the resulting 615 samples and 2,366,788 SNPs to standard QC measures. An identity-by-descent (IBD) check identified 90 sample duplicates and 1 sample pair with a sibling or child/parent relation. Each of these related pairs involved a WGS sample and an imputed sample (i.e., the duplicates were the result of each sample being represented in both the imputed and WGS datasets). For these samples, we kept the data from the WGS dataset as it appeared to be of slightly better quality overall, and we dropped the imputed duplicates. No additional variants required removal. Note that because imputed data were not available for the trisomic chromosome 21 (methods for imputing trisomic genotypes are lacking), this target dataset for the primary analyses did not include chromosome 21 variants. This intermediate data set included 524 samples (263 cases, 261 controls) and 2,366,788 autosomal SNPs.

We then performed PCA, first anchoring our dataset in the HapMap3 (International HapMap Consortium 2005) dataset and constructing PCs (using PLINK1.9) to identify and remove DS samples with PC values outside of the HapMap3 CEU cluster (in order to match the European ancestry of the discovery datasets), and

then removing the HapMap samples and performing further outlier removal based only on the DS samples (see Supplemental Methods for details). This PCA process identified 37 sample outliers for removal.

As a final step in preparing the DS target dataset for PRS analysis, we removed the major histocompatibility complex region (Chr6: 25-34 Mb, hg19), which is a region of extended high linkage disequilibrium that can overly influence PRS results. Our final data set included 487 samples (245 DS+AVSD cases, 242 DS+NH controls) and 2,351,951 autosomal SNPs (excluding chromosome 21). The multiple steps involved in generating this final data set for the primary PRS analyses are presented as a flowchart in Supplemental Figure 2.

Target dataset for secondary PRS analyses

Our secondary PRS analyses examined the contribution by alleles on the trisomic chromosome 21 to a polygenic component for DS-associated AVSD. We were able to do this because, while imputed data were not available for chromosome 21, all imputed samples did have SNP array genotypes for chromosome 21. Furthermore, the WGS samples had sequencing data for chromosome 21. For all target samples analyzed in the primary analyses (245 cases, 242 controls), we therefore obtained SNP array data for the trisomic chromosome 21, and likewise limited all other chromosomes to SNPs available on the Affymetrix Genome-Wide Human SNP 6.0 array.

We processed the chromosome 21 data separately from the other chromosomes due to the trisomic nature of these data. Prior to merging chromosome 21 data for the

imputed and WGS samples, we applied certain QC filters (see Supplemental Methods for details). We subsequently merged the array and WGS chromosome 21 data, leaving 3,984 chromosome 21 SNPs and 487 samples.

Discovery data used to define weights for the PRS

For discovery datasets, there were no GWAS of AVSD or other congenital heart defects (CHD) among individuals with DS that were independent of our target dataset nor were there any GWAS specifically for non-syndromic AVSD. Thus, we used results from two of the largest available independent GWAS of mixed CHD, diagnosed among those without DS who were ancestrally matched to our target samples.

The first discovery dataset was a GWAS of 2,594 cases with a mixture of CHD diagnoses (see Table 3.3) and 5,159 population-based controls, all of European ancestry. Genotyping was performed using the Illumina Human660W-Quad array for cases and the Illumina 1.2M chip for controls. The GWAS results included summary statistics for 501,899 autosomal SNPs. Summary level results for this GWAS are available upon request through Dr. Heather Cordell. GWAS of particular diagnostic CHD subsets of this dataset have been published previously (Cordell, Töpf, et al. 2013; Cordell, Bentham, et al. 2013).

The second discovery dataset was a GWAS of 406 mixed CHD cases (Table 3.4) and 2,976 pediatric controls, all recruited from the same hospital and self-reporting as non-Hispanic Caucasian (Agopian et al. 2017). Samples were genotyped with Illumina arrays (550 v1/v3, 610, or 2.5M chip), and genome-wide imputation was then carried

out using the 1000 Genomes Project data as a reference. The GWAS results included summary statistics for 4,612,359 autosomal SNPs, all of which had imputation $r_2 > 0.80$. Summary results from this GWAS are available upon request through Dr. A.J. Agopian.

We used each of these discovery datasets separately as training data for the PRS analyses. We also meta-analyzed the summary results from these two GWAS using GWAMA (Mägi and Morris 2010), and used the resulting estimates as training data.

Generating PRS for the primary analyses

For the primary PRS analyses, PRSice-2 (Choi and O'Reilly 2019) (version 2.1.6) was used to generate PRS for each sample in the target dataset. Prior to PRS construction, PRSice performs clumping on the discovery dataset in order to obtain a set of independent SNPs for scoring (clumping parameters: 500kb window, r_2 threshold 0.10). The clumped SNPs are then used to generate PRS, which are calculated as

$$PRS_j = \sum_i \frac{\beta_i \times EA_{ij}}{N_j}$$

where the subscript *i* denotes a specific SNP contributing to the PRS, the subscript *j* denotes a particular individual in the target dataset, *β* is the estimated effect from the discovery GWAS (e.g., the natural logarithm of the odds ratio), *EA* is the number of effective alleles possessed by the target individual (0,1 or 2 for a disomic chromosome),

and *N* is the total number of alleles considered for scoring. To facilitate interpretation of results, we applied an option in PRSice to standardize the PRS.

We constructed multiple PRS for each target individual using different subsets of the set of clumped SNPs, with subsets determined by applying different p-value thresholds to the discovery GWAS results (e.g., PRS may be constructed using SNPs with discovery p-value < $1x10^{-6}$, < 0.05, < 1). Given the relatively small sample sizes for each discovery GWAS, we were concerned that effect estimates for SNPs with lower MAFs may be particularly subject to error. To address this, we applied a range of MAF filters (from 0.10 to 0.40) to the discovery datasets prior to generating the PRS, excluding those SNPs with MAF below the threshold. Thus, for each discovery dataset, we constructed PRS and performed separate analyses for each combination of MAF filter and discovery p-value threshold.

Generating PRS for the secondary analyses

For the secondary PRS analyses, which involved analyses both with and without the trisomic chromosome 21 data, we constructed PRS using PLINK1.9. The PLINK1.9 binary, which is the file format that we used in conjunction with PRSice for the primary PRS analyses, is not able to represent trisomic genotype data. However, we were able to modify the chromosome 21 genotype data to fit the PLINK1.9 dosage file format, which can be used in conjunction with PLINK's allelic scoring flag to generate PRS. This involved dividing each allele count by 3 and thereby converting allele counts of 0, 1, 2 and 3 to values of 0, 1/3, 2/3 and 1 (interpreted by PLINK as dosages ranging from 0 to

1). We then used this chromosome 21 dosage format file in combination with the clumped training data (clumped using PRSice) to generate PRS, which were generated by PLINK as a simple sum score (a sum of the products of SNP weight times transformed allele count for each scoring SNP). Finally, we multiplied each outputted PRS by 3, yielding PRS that accurately reflected allele counts of 0, 1, 2 and 3 for the trisomic chromosome 21.

Separately, we used PLINK1.9 to construct PRS for the remaining autosomes. Given that these remaining autosomes were diploid, we were able to use the standard PLINK1.9 binary in combination with the allelic scoring flag to generate PRS. To be consistent with the chromosome 21 PRS, we used an option to generate these PRS as sum scores. For the analyses including chromosome 21, we then summed the chromosome 21 PRS and the PRS for the remaining autosomes for each target individual, yielding a PRS based on alleles from all autosomes combined. The analyses excluding chromosome 21 only utilized the PRS based on all autosomes minus chromosome 21. As for the primary PRS analyses, we standardized the final PRS, and generated multiple PRS for each target individual based on different discovery GWAS p-value and MAF thresholds.

Testing association of PRS with DS+AVSD

We used logistic regression to test associations of PRS with the outcome; this was performed by PRSice for the primary analyses and within R for the secondary analyses. We included sex, platform (WGS vs. imputed), and the top five principal

components of ancestry as covariates in the analyses. Given the multiple testing involved in these PRS analyses (394 tests for different combinations of MAF filter, p-value threshold, and discovery and target datasets, considering the primary and secondary PRS analyses together), we employed the p-ACT (Conneely and Boehnke 2007) method to generate p-values corrected for multiple correlated tests.

RESULTS

Gene discovery using SKAT analyses

Three separate SKAT-O analyses were conducted at different MAF filtering thresholds. All of the p-values for SKAT-O analyses derived from the WES dataset fell within expected or slightly deflated values, likely due to the small sample size (Figure 3.1). No gene was significant following Bonferroni correction for the total number of genes in each set, although 19 genes in the common variant analysis (MAF > 0.05), 10 genes in the rare variant analysis (MAF < 0.01 or missing in gnomAD) and one gene in the ultra-rare variant analysis (MAF < 0.001) displayed nominal significance levels of p-value < 10^{-3} (Tables 3.5-3.7). Of those genes with nominal significance in the WES dataset, three were supported in the WGS dataset. Two of those genes, *NOTCH4* and *CEP290*, have been reported as being involved in heart development. *NOTCH4* is expressed in the developing heart and has previously been identified as playing a role in early artery and endothelial-to-mesenchymal transformation, which is critical for endocardial cushion differentiation (Noseda et al. 2004; Wythe et al. 2013). *CEP290*

codes for a centrosomal protein involved in cilia development that has been found to have differential expression between the left and right ventricles of the heart in newborn piglets, and which may play a role in remodeling of the ventricular myocardium postnatally (Torrado et al. 2010). The third gene, ZNF318, has not previously been implicated in heart defects.

As a follow-up of these results, we conducted gene-set analyses based on genes in the ciliome (van Dam et al. 2013) and those in the Notch pathway (Kanehisa 2000) using the WES dataset. We used SKAT-O, combining all variants identified in each pathway and again filtering on MAF thresholds, and found moderate significance of $p < 0.05$ in the set of rare variants (MAF < 0.01 or missing in gnomAD) (Table 3.8).

CHD polygenic risk score and its association with DS+AVSD

Primary analyses indicate a non-significant association of the CHD-based PRS with DS+AVSD

Over a range of MAF filters and discovery GWAS p-value thresholds for constructing PRS, the analyses using the GWAS of 2,594 mixed CHD cases and 5,159 controls as the discovery dataset (501,899 autosomal SNPs) tended to yield maximum odds ratios (ORs) of 1.2 to 1.3 for association of PRS with AVSD among those with DS, meaning that a 1 standard deviation (SD) increase in PRS was associated with a 20-30% greater odds of having AVSD in the DS target sample (Figure 3.4). Corresponding Nagelkerke's r_2 values ranged from 0.75-1.25% (calculated as Nagelkerke's r_2 for the model with PRS and covariates minus Nagelkerke's r_2 for the model with only

covariates), with p-values that were non-significant following adjustment for multiple correlated tests (adjusted p-values > 0.15; unadjusted p-values approximately 0.01-0.09). These maximum results were most evident at higher MAF filters (i.e., MAF ≥ 0.30, ≥ 0.35, ≥ 0.40) and discovery GWAS p-value thresholds between 0.001 and 0.3. Figure 3.2 and Table 3.9 present results when PRS are constructed using SNPs with MAF ≥ 0.35, which are representative of the maximum PRS results achieved when using this particular discovery dataset.

PRS results when using the GWAS of 406 CHD cases and 2,976 pediatric controls as the discovery dataset (4,612,359 autosomal SNPs) exhibited a different pattern than when using the GWAS of 2,594 mixed CHD cases and 5,159 controls as training data. Across various MAF filters and p-value thresholds, ORs tended to hover near the null and on both sides of the null, indicating that these PRS were minimally associated with AVSD (Figure 3.5). A few results were stronger, with ORs in the 1.2 to 1.3 range (adjusted p-values > 0.15); these results occurred when using MAF filters of ≥ 0.10 and ≥ 0.15 in combination with the smallest discovery GWAS p-value thresholds for selecting scoring SNPs.

We also performed a meta-analysis of the two GWAS datasets, yielding a single discovery dataset with association estimates for 4,684,854 autosomal SNPs, of which 429,336 SNPs had estimates based on both studies (meta-analysis sample size of 3,000 CHD cases and 8,135 controls), while the remainder had estimates based on just one of the two studies. In constructing PRS based on this meta-analysis discovery dataset, we applied an inverse variance weighting approach such that SNP association estimates based on a larger sample size (e.g., two studies) were weighted more

heavily. Using the meta-analysis dataset in this manner produced results which, as might be expected, were a mixture of the PRS results obtained when using each discovery GWAS dataset separately (Figure 3.6). In general, maximum ORs for association of AVSD in DS with PRS and corresponding Nagelkerke's r_2 values were slightly attenuated compared with results when using the GWAS of 2,594 mixed CHD cases and 5,159 controls as the discovery dataset.

Adding data from chromosome 21 into the PRS calculation did not change the association with DS+AVSD

We performed the secondary analyses using only the training dataset derived from 2,594 mixed CHD cases and 5,159 controls, since using these training data produced the best results for the primary PRS analyses. The results from PRS analyses including and excluding chromosome 21 were essentially the same, with only slight fluctuations in ORs and corresponding Nagelkerke's r_2 values (Figures 3.7 and 3.8). These results generally followed a similar pattern to those observed for the primary PRS analysis using the same discovery dataset (Figure 3.4), wherein use of greater MAF filters yielded larger associations. However, the results from these secondary analyses fluctuated more across discovery GWAS p-value thresholds and included more outlier OR estimates, which was likely a result of the smaller number of SNPs used for scoring in the secondary analyses (which were limited to SNPs on the Affymetrix array).

DISCUSSION

Previous studies of AVSD in DS have had limited success in identifying rare variant contributions and have failed to clarify the role of common variants (Sailani et al. 2013; Ramachandran, Zeng, et al. 2015). In the current study, we examined the role of rare and common SNVs in DS-associated AVSD by analyzing data from whole exome sequencing, whole genome sequencing, and genome-wide SNP imputation in cases with DS+AVSD and DS+NH controls. We used SNV-set analyses (grouping variants into genes and pathways) to examine both rare and common variant associations and polygenic risk score methods to investigate the combined effect of common variants across the genome.

In the genome-wide variant-set analyses which grouped SNVs by gene, we obtained preliminary support for 3 genes, 2 of which have been reported previously as genes involved in heart development. Prior studies in AVSD and other heart defects have identified cilia as a major factor in heart development (Burnicka-Turek et al. 2016; Klena, Gibbs, and Lo 2017) and the ciliome has been identified as a pathway enriched in DS+AVSD for rare deletions and differential gene expression (Ripoll et al. 2012; Ramachandran, Zeng, et al. 2015). In our analyses, *NOTCH4* and *CEP290*, whose roles in heart development and the ciliome have been previously described (Noseda et al. 2004; Torrado et al. 2010), were found to have a nominally significant enrichment among DS+AVSD cases in both WES and WGS datasets. *CEP290* was also recently identified as potentially associated with non-syndromic CHD (including any type of heart defect) by a targeted sequencing study of 406 candidate genes involved in heart development (Alharbi et al. 2018).

Investigating rare variants in the Notch pathway and ciliome genes as a whole also displayed a moderate enrichment in our dataset. These results provide further support for the involvement of the Notch pathway and ciliome in heart development, and suggest a specific link between these pathways and AVSD in DS. Considering the small sample sizes of the WES and WGS datasets and the case-control imbalance in the WGS dataset, a larger balanced WGS dataset should enable greater power to detect individual genes in these pathways that contribute to AVSD in DS.

The PRS analyses are the first such analyses of AVSD in DS, and to the best of our knowledge they are also the first use of PRS methods to examine polygenicity of CHD generally. Our analyses of PRS calculated from GWAS studies of non-syndromic CHD suggest at minimum a small polygenic contribution to AVSD among individuals with DS. When using dense SNP data (WGS or imputed data) for the 487 individuals in the target sample and excluding chromosome 21, a single standard deviation increase in PRS was associated with a 20-30% increased odds for having AVSD, with Nagelkerke's r_2 values for PRS of around 1% (Figure 3.2; Figure 3.4); this occurred when using the larger of the two independent discovery datasets. Assuming a population prevalence of 20% for AVSD among those with DS, these Nagelkerke's r_2 values are quite similar to the corresponding liability scale r_2 values (correcting for case-control ascertainment). For instance, the PRS analyses depicted in Figure 3.2 yielded a Nagelkerke's r_2 of 1.03% when applying MAF ≥ 0.35 and discovery GWAS p-value ≤ 0.001 thresholds; the corresponding liability r_2 estimate is 1.11% (S. H. Lee et al. 2012). As demonstrated by the PRS results presented in Figures 3.7 and 3.8, which involved the use of array SNPs only, inclusion of dense genotype data for chromosome 21 is

unlikely to substantially alter these estimates for the association of PRS with DS-associated AVSD; SNPs on chromosome 21 are perhaps not a key factor driving AVSD in DS.

Given the small sample sizes for the discovery GWAS datasets and prior research demonstrating that variance explained by PRS tends to increase as discovery GWAS sample size increases (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014), which is attributable to increased accuracy of the SNP effect estimates used as weights for the PRS, it seems likely that the use of a larger discovery GWAS of CHD will uncover a greater polygenic contribution to AVSD in DS. Furthermore, use of a large discovery GWAS that only includes the particular CHD subtypes which are most closely genetically related to AVSD (perhaps a GWAS including only AVSD and septal defect cases) may reveal a polygenic contribution to DS-associated AVSD that exceeds what we have identified. We demonstrate this in Figure 3.9, showing that under reasonable assumptions, using a discovery GWAS of phenotypes that are highly genetically correlated with the target phenotype (AVSD) will result in PRS r_2 values that increase as discovery GWAS sample size increases; the discovery samples similar in size to those used for the current PRS analyses are only able to capture a portion of the true polygenic component (plots generated using the 'avengeme' R package; Dudbridge 2013).

The finding of an association of AVSD in DS with PRS constructed based on SNPs identified as having some measure of association with CHD in mixed CHD samples suggests the possibility of genetic overlap between AVSD and various other subtypes of CHD. This is consistent with the potential for investigating DS-associated

AVSD to shed light on fundamental biology relevant to CHD more generally. To further examine this potential genetic overlap, including which CHD subtypes may have the greatest shared genetic architecture with AVSD, it will be important to utilize large GWAS datasets of specific CHD subtypes rather than a mixture of CHD types.

We observed that PRS constructed based on the discovery GWAS of 2,594 mixed CHD cases and 5,159 controls consistently yielded ORs > 1 (indicating, as expected, that increased PRS was associated with increased AVSD risk). In contrast, PRS constructed using the discovery GWAS of 406 CHD cases and 2,976 pediatric controls yielded OR estimates generally quite close to the null, and on both sides of the null. One possible reason for this difference is that the smaller-sized discovery GWAS had more imprecisely estimated SNP associations, leading to less informative PRS. Another possibility is that particular CHD diagnoses included within the larger discovery GWAS may be more genetically related to AVSD in DS than the CHD diagnoses in the smaller discovery GWAS. Indeed, the larger GWAS included 73 cases with AVSD, while in the smaller GWAS there were only seven instances of AVSD (six of the cases with double outlet right ventricle also had AVSD, and a single case had tetralogy of Fallot with atrioventricular canal septal defect).

In conclusion, while our analyses yielded no statistically significant findings following multiple testing correction, the results suggest that rare variation in certain pathways and common variants acting through a polygenic component may play roles in increasing risk for AVSD among those with DS. The use of larger sample sizes, including a larger DS+AVSD/DS+NH sample as well as larger discovery GWAS samples for PRS construction, is important as it will enable greater power for identifying

and quantifying rare variant and polygenic contributions. It is also possible that genetic effects on DS-associated AVSD are particularly pronounced in the presence of certain environmental factors. This could be investigated in future studies by examining environmental interactions with potentially involved genetic factors including variation in the Notch pathway and ciliome as well as PRS.

Availability of data:

Affymetrix Genome-Wide Human SNP 6.0 array genotype data are available for 437 DS samples (DS+AVSD cases and DS+NH controls) via the Gene Expression Omnibus [GEO] data repository, accession number GSE60607. Genotypes for some of the samples described in this paper were excluded from GEO due to privacy concerns.

WES data is available on request.

WGS data results from the PCGC samples can be accessed through the PCGC dbGaP study (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001194.v2.p2).

References

Ackerman, Christine, Adam E. Locke, Eleanor Feingold, Benjamin Reshey, Karina Espana, Janita Thusberg, Sean Mooney, et al. 2012. “An Excess of Deleterious Variants in VEGF-A Pathway Genes in Down-Syndrome-Associated Atrioventricular Septal Defects.” *American Journal of Human Genetics* 91 (4): 646–59. https://doi.org/10.1016/j.ajhg.2012.08.017.

Agopian, A. J., Elizabeth Goldmuntz, Hakon Hakonarson, Anshuman Sewda, Deanne Taylor, Laura E. Mitchell, and Pediatric Cardiac Genomics Consortium*. 2017. “Genome-Wide Association Studies and Meta-Analyses for Congenital Heart Defects.” *Circulation. Cardiovascular Genetics* 10 (3): e001449. https://doi.org/10.1161/CIRCGENETICS.116.001449.

Al Turki, Saeed, Ashok K. Manickaraj, Catherine L. Mercer, Sebastian S. Gerety, Marc-Phillip Hitz, Sarah Lindsay, Lisa C. A. D’Alessandro, et al. 2014. “Rare Variants in NR2F2 Cause Congenital Heart Defects in Humans.” *American Journal of Human Genetics* 94 (4): 574–85. https://doi.org/10.1016/j.ajhg.2014.03.007.

Alharbi, Khalid M., Abdelhadi H. Al-Mazroea, Atiyeh M. Abdallah, Yousef Almohammadi, S. Justin Carlus, and Sulman Basit. 2018. “Targeted Next-Generation Sequencing of 406 Genes Identified Genetic Defects Underlying Congenital Heart Disease in Down Syndrome Patients.” *Pediatric Cardiology* 39 (8): 1676–80. https://doi.org/10.1007/s00246-018-1951-3.

Anderson, Carl A., Fredrik H. Pettersson, Geraldine M. Clarke, Lon R. Cardon, Andrew P. Morris, and Krina T. Zondervan. 2010. “Data Quality Control in Genetic Case-

Control Association Studies." *Nature Protocols* 5 (9): 1564–73. https://doi.org/10.1038/nprot.2010.116.

Burnicka-Turek, Ozanna, Jeffrey D. Steimle, Wenhui Huang, Lindsay Felker, Anna Kamp, Junghun Kweon, Michael Peterson, et al. 2016. "Cilia Gene Mutations Cause Atrioventricular Septal Defects by Multiple Mechanisms." *Human Molecular Genetics* 25 (14): 3011–28. https://doi.org/10.1093/hmg/ddw155.

Chang, Christopher C., Carson C. Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4: 7. https://doi.org/10.1186/s13742-015-0047-8.

Chen, Vincent C., Robert Stull, Daniel Joo, Xin Cheng, and Gordon Keller. 2008. "Notch Signaling Respecifies the Hemangioblast to a Cardiac Fate." *Nature Biotechnology* 26 (10): 1169–78. https://doi.org/10.1038/nbt.1497.

Choi, Shing Wan, and Paul F. O'Reilly. 2019. "PRSice-2: Polygenic Risk Score Software for Biobank-Scale Data." *GigaScience* 8 (7). https://doi.org/10.1093/gigascience/giz082.

Conneely, Karen N., and Michael Boehnke. 2007. "So Many Correlated Tests, so Little Time! Rapid Adjustment of P Values for Multiple Correlated Tests." *American Journal of Human Genetics* 81 (6): 1158–68. https://doi.org/10.1086/522036.

Cordell, Heather J., Jamie Bentham, Ana Topf, Diana Zelenika, Simon Heath, Chrysovalanto Mamasoula, Catherine Cosgrove, et al. 2013. "Genome-Wide Association Study of Multiple Congenital Heart Disease Phenotypes Identifies a Susceptibility Locus for Atrial Septal Defect at Chromosome 4p16." *Nature*

Genetics 45 (7): 822–24. https://doi.org/10.1038/ng.2637.

Cordell, Heather J., Ana Töpf, Chrysovalanto Mamasoula, Alex V. Postma, Jamie Bentham, Diana Zelenika, Simon Heath, et al. 2013. “Genome-Wide Association Study Identifies Loci on 12q24 and 13q32 Associated with Tetralogy of Fallot.” *Human Molecular Genetics* 22 (7): 1473–81. https://doi.org/10.1093/hmg/dds552.

Dam, Teunis JP van, Gabrielle Wheway, Gisela G Slaats, Martijn A Huynen, and Rachel H Giles. 2013. “The SYSCILIA Gold Standard (SCGSv1) of Known Ciliary Components and Its Applications within a Systems Biology Consortium.” *Cilia* 2 (May): 7. https://doi.org/10.1186/2046-2530-2-7.

Das, Sayantan, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E. Locke, Alan Kwong, Scott I. Vrieze, et al. 2016. “Next-Generation Genotype Imputation Service and Methods.” *Nature Genetics* 48 (10): 1284–87. https://doi.org/10.1038/ng.3656.

Dudbridge, Frank. 2013. “Correction: Power and Predictive Accuracy of Polygenic Risk Scores.” *PLOS Genetics* 9 (4): 10.1371/annotation/b91ba224. https://doi.org/10.1371/annotation/b91ba224-10be-409d-93f4-7423d502cba0.

Ferencz, C., C. A. Neill, J. A. Boughman, J. D. Rubin, J. I. Brenner, and L. W. Perry. 1989. “Congenital Cardiovascular Malformations Associated with Chromosome Abnormalities: An Epidemiologic Study.” *The Journal of Pediatrics* 114 (1): 79–86. https://doi.org/10.1016/s0022-3476(89)80605-5.

Hartman, Robert J., Tiffany Riehle-Colarusso, Angela Lin, Jaime L. Frías, Sonali S. Patel, Kara Duwe, Adolfo Correa, Sonja A. Rasmussen, and National Birth Defects Prevention Study. 2011. “Descriptive Study of Nonsyndromic

Atrioventricular Septal Defects in the National Birth Defects Prevention Study, 1997-2005." *American Journal of Medical Genetics. Part A* 155A (3): 555–64. https://doi.org/10.1002/ajmg.a.33874.

Hoang, Thanh T., Elizabeth Goldmuntz, Amy E. Roberts, Wendy K. Chung, Jennie K. Kline, John E. Deanfield, Alessandro Giardini, et al. 2018. "The Congenital Heart Disease Genetic Network Study: Cohort Description." *PloS One* 13 (1): e0191319. https://doi.org/10.1371/journal.pone.0191319.

International HapMap Consortium. 2005. "A Haplotype Map of the Human Genome." *Nature* 437 (7063): 1299–1320. https://doi.org/10.1038/nature04226.

Ionita-Laza, Iuliana, Seunggeun Lee, Vlad Makarov, Joseph D. Buxbaum, and Xihong Lin. 2013. "Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants." *American Journal of Human Genetics* 92 (6): 841–53. https://doi.org/10.1016/j.ajhg.2013.04.015.

Johnston, H. Richard, Pankaj Chopra, Thomas S. Wingo, Viren Patel, International Consortium on Brain and Behavior in 22q11.2 Deletion Syndrome, Michael P. Epstein, Jennifer G. Mulle, Stephen T. Warren, Michael E. Zwick, and David J. Cutler. 2017. "PEMapper and PECaller Provide a Simplified Approach to Whole-Genome Sequencing." *Proceedings of the National Academy of Sciences of the United States of America* 114 (10): E1923–32. https://doi.org/10.1073/pnas.1618065114.

Kanehisa, Minoru. 2000. *Post-Genome Informatics*. Oxford University Press.

Khera, Amit V., Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, et al. 2018. "Genome-Wide Polygenic

Scores for Common Diseases Identify Individuals with Risk Equivalent to Monogenic Mutations." *Nature Genetics* 50 (9): 1219–24. https://doi.org/10.1038/s41588-018-0183-z.

Klena, Nikolai T., Brian C. Gibbs, and Cecilia W. Lo. 2017. "Cilia and Ciliopathies in Congenital Heart Disease." *Cold Spring Harbor Perspectives in Biology* 9 (8). https://doi.org/10.1101/cshperspect.a028266.

Kotlar, Alex V., Cristina E. Trevino, Michael E. Zwick, David J. Cutler, and Thomas S. Wingo. 2018. "Bystro: Rapid Online Variant Annotation and Natural-Language Filtering at Whole-Genome Scale." *Genome Biology* 19 (1): 14. https://doi.org/10.1186/s13059-018-1387-3.

Lee, Seunggeun, Mary J. Emond, Michael J. Bamshad, Kathleen C. Barnes, Mark J. Rieder, Deborah A. Nickerson, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, David C. Christiani, Mark M. Wurfel, and Xihong Lin. 2012. "Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies." *American Journal of Human Genetics* 91 (2): 224–37. https://doi.org/10.1016/j.ajhg.2012.06.007.

Lin, Fu-Jung, Li-Ru You, Cheng-Tai Yu, Wen-Hsin Hsu, Ming-Jer Tsai, and Sophia Y. Tsai. 2012. "Endocardial Cushion Morphogenesis and Coronary Vessel Development Require Chicken Ovalbumin Upstream Promoter-Transcription Factor II." *Arteriosclerosis, Thrombosis, and Vascular Biology* 32 (11): e135–46. https://doi.org/10.1161/ATVBAHA.112.300255.

Mägi, Reedik, and Andrew P. Morris. 2010. "GWAMA: Software for Genome-Wide

Association Meta-Analysis." *BMC Bioinformatics* 11 (May): 288. https://doi.org/10.1186/1471-2105-11-288.

McCarthy, Shane, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R. Wood, Alexander Teumer, Hyun Min Kang, et al. 2016. "A Reference Panel of 64,976 Haplotypes for Genotype Imputation." *Nature Genetics* 48 (10): 1279–83. https://doi.org/10.1038/ng.3643.

Noseda, Michela, Graeme McLean, Kyle Niessen, Linda Chang, Ingrid Pollet, Rachel Montpetit, Réza Shahidi, et al. 2004. "Notch Activation Results in Phenotypic and Functional Changes Consistent with Endothelial-to-Mesenchymal Transformation." *Circulation Research* 94 (7): 910–17. https://doi.org/10.1161/01.RES.0000124300.76171.C9.

Pereira, Fred A., Yuhong Qiu, Ge Zhou, Ming-Jer Tsai, and Sophia Y. Tsai. 1999. "The Orphan Nuclear Receptor COUP-TFII Is Required for Angiogenesis and Heart Development." *Genes & Development* 13 (8): 1037–49.

Purcell, Shaun M., and Christopher C. Chang. n.d. *PLINK [v1.9b6.6]*. www.cog-genomics.org/plink/1.9.

R: A Language and Environment for Statistical Computing. 2017. Vienna, Austria: R Foundation for Statistical Computing.

Ramachandran, Dhanya, Jennifer G. Mulle, Adam E. Locke, Lora J. H. Bean, Tracie C. Rosser, Promita Bose, Kenneth J. Dooley, et al. 2015. "Contribution of Copy-Number Variation to Down Syndrome-Associated Atrioventricular Septal Defects." *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 17 (7): 554–60. https://doi.org/10.1038/gim.2014.144.

Ramachandran, Dhanya, Zhen Zeng, Adam E. Locke, Jennifer G. Mulle, Lora J. H. Bean, Tracie C. Rosser, Kenneth J. Dooley, et al. 2015. “Genome-Wide Association Study of Down Syndrome-Associated Atrioventricular Septal Defects.” *G3 (Bethesda, Md.)* 5 (10): 1961–71. https://doi.org/10.1534/g3.115.019943.

Rambo-Martin, Benjamin L., Jennifer G. Mulle, David J. Cutler, Lora J. H. Bean, Tracie C. Rosser, Kenneth J. Dooley, Clifford Cua, et al. 2018. “Analysis of Copy Number Variants on Chromosome 21 in Down Syndrome-Associated Congenital Heart Defects.” *G3 (Bethesda, Md.)* 8 (1): 105–11. https://doi.org/10.1534/g3.117.300366.

Rayner, W. n.d. *Script to Check Plink .Bim Files against HRC/1000G for Strand, Id Names, Positions, Alleles, Ref/Alt Assignment [v 4.2.9].* https://www.well.ox.ac.uk/~wrayner/tools/.

Ripoll, Clémentine, Isabelle Rivals, Emilie Ait Yahya-Graison, Luce Dauphinot, Evelyne Paly, Clothilde Mircher, Aimé Ravel, et al. 2012. “Molecular Signatures of Cardiac Defects in Down Syndrome Lymphoblastoid Cell Lines Suggest Altered Ciliome and Hedgehog Pathways.” *PLoS ONE* 7 (8). https://doi.org/10.1371/journal.pone.0041616.

Sailani, M. Reza, Periklis Makrythanasis, Armand Valsesia, Federico A. Santoni, Samuel Deutsch, Konstantin Popadin, Christelle Borel, et al. 2013. “The Complex SNP and CNV Genetic Architecture of the Increased Risk of Congenital Heart Defects in Down Syndrome.” *Genome Research* 23 (9): 1410–21. https://doi.org/10.1101/gr.147991.112.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014. "Biological Insights from 108 Schizophrenia-Associated Genetic Loci." *Nature* 511 (7510): 421–27. https://doi.org/10.1038/nature13595.

"The Congenital Heart Disease Genetic Network Study." 2013. *Circulation Research* 112 (4): 698–706. https://doi.org/10.1161/CIRCRESAHA.111.300297.

Torrado, Mario, Raquel Iglesias, Beatriz Nespereira, and Alexander T. Mikhailov. 2010. "Identification of Candidate Genes Potentially Relevant to Chamber-Specific Remodeling in Postnatal Ventricular Myocardium." *Journal of Biomedicine and Biotechnology* 2010. https://doi.org/10.1155/2010/603159.

Supplemental Methods

Target dataset for primary PRS analysis

Imputed samples:

There originally were 459 DS samples (211 DS+AVSD cases, 248 DS+NH controls) with Affymetrix Genome-Wide Human SNP 6.0 array genotype data, including the 210 cases and 242 controls analyzed in the prior GWAS of DS-associated AVSD (Ramachandran et al. 2015). Using PLINK1.9 (version 1.90b6.6)(Chang et al. 2015; Purcell and Chang, n.d.) and R (version 3.4.1)(*R: A Language and Environment for Statistical Computing* 2017), we applied standard GWAS QC procedures, excluding subjects for sex discordance, outlier heterozygosity rates (+/- 3 SDs from the mean), missing > 3% of genotypes, and one subject from each pair having IBD > 0.1875. Variant filters included missing for > 5% of samples, MAF < 0.01, HWE mid-p-value < 0.00001 (among controls), and significantly different rates of missingness in cases versus controls ($p < 0.00001$). We then used principal component analysis (PCA) to identify and remove any population outliers, which involved identifying and removing non-European samples using the HapMap3 (International HapMap Consortium 2005) dataset as a population reference (we identified ancestral outliers based on the Anderson et al. 2010 protocol). All together, these QC steps yielded a dataset with 207 DS+AVSD cases and 234 DS+NH controls, and 612,125 autosomal SNPs (excluding chromosome 21).

For these samples, we then performed genotype imputation using the Michigan Imputation Server (Das et al. 2016). Prior to imputation, all alleles were aligned to the (+) strand, and we used a program (Rayner) written by the McCarthy Group to check our dataset against the Haplotype Reference Consortium (HRC) panel and ensure that our data were properly configured for imputation using the HRC panel. We then submitted the DS dataset to the Michigan Imputation Server, for imputation based on the HRC panel (version r1-1 2016)(McCarthy et al. 2016), which includes 32,470 samples predominantly of European ancestry.

The post-imputation files included 38,596,402 autosomal variants (all SNPs). Mean correlation between true and imputed genotypes for the ~600,000 genotyped SNPs was 0.990, suggesting high quality imputation. Considering all post-imputation variants, those with MAF ≥ 0.05 (5,349,403 variants) had mean imputation r_2 = 0.971, those with 0.01 ≤ MAF < 0.05 (2,300,344 variants) had mean r_2 = 0.882, and those with MAF < 0.01 (30,946,655 variants) had mean r_2 = 0.180. This indicates good imputation quality for variants with common or moderate MAF. We decided to drop variants with MAF < 0.01, those missing for more than 2% of samples, those with a maximum imputed genotype probability < 0.80, and those with imputation r_2 < 0.80.

We then applied standard GWAS QC to the imputed dataset. We dropped one sample with an outlying heterozygosity rate (> 3 SDs below the mean). No samples were dropped for excess missing genotypes (all had < 1% missingness). Following removal of the single sample, we again excluded variants missing for > 2% of individuals and those with MAF < 0.01, and also dropped variants with HWE mid-p-value < 0.00001 and those with significant differences in missing genotype rate between

cases and controls ($p < 0.05$). We also removed variants with A/T, T/A, C/G, and G/C alleles which can be difficult to match between datasets due to strand ambiguity. This left a dataset with 440 samples (206 DS+AVSD cases, 234 DS+NH controls) and 5,079,537 autosomal SNPs.

WGS samples:

Starting from the previously described post-QC WGS dataset, which included 175 samples (148 DS+AVSD cases, 27 DS+NH controls), we applied additional variant filters in order to more closely match the variant QC procedures which had been applied to the imputed dataset. We removed variants with MAF < 0.01, those missing for > 2% of samples, and indels, leaving a WGS dataset with 175 samples and 4,173,676 autosomal SNPs (excluding chromosome 21).

Merging WGS and imputed samples:

Coordinates for the WGS dataset were based on hg38, while those for the imputed dataset were based on hg19. Prior to merging the datasets, we used the UCSC Genome Browser (Kent et al. 2002) LiftOver tool to convert the WGS data coordinates from hg38 to hg19, and also modified rsIDs as needed using an external file based on HRC panel variants containing hg19 rsIDs and coordinates. We chose to convert the WGS data to hg19 rather than converting the imputed data to hg38 as a matter of convenience, given the PRS training files we used had hg19 coordinates.

As one additional step prior to merging the WGS and imputed datasets, we compared allele frequencies for SNPs in each dataset in order to identify any instances

where allele frequency for a SNP in one dataset differed significantly from its allele frequency in the other dataset, which could indicate genotyping error for the variant. We identified and removed 77 SNPs with allele frequencies that differed by at least 0.20 between the WGS and imputed datasets.

We then merged the WGS and imputed datasets on rsID, position, and alleles (using PLINK1.9), yielding a single dataset with 615 samples and 2,366,788 SNPs. For all 615 samples missingness was < 1%. An identity-by-descent (IBD) check identified 90 sample duplicates and 1 sample pair with a sibling or child/parent relation. Each of these related pairs involved a WGS sample and an imputed sample (i.e., the duplicates were the result of each sample being represented in both the imputed and WGS datasets). For these samples, we kept the data from the WGS dataset as it appeared to be of slightly better quality overall, and we dropped the imputed duplicates. No additional variant QC filters were needed -- all SNPs had missingness ≤ 2% among all samples and ≤ 3% among both cases and controls, all had MAF approximately ≥ 1% (we applied stricter MAF filters during PRS construction), and no SNPs required dropping for HWE violation. Thus, this intermediate data set included 524 samples (263 cases, 261 controls) and 2,366,788 autosomal SNPs.

We next performed PCA, first anchoring our dataset in the HapMap3 dataset and constructing PCs to identify and remove DS samples with PC values outside of the HapMap3 CEU cluster (in order to match the European ancestry of the discovery datasets), and then removing the HapMap samples and performing further outlier removal based only on the DS samples. We constructed PCs for just the DS samples, and removed samples with values > 3 SD from the mean for PC1 or PC2 (which

explained most of the genetic variation in the sample). We then reconstructed PCs for the remaining samples and again identified 3 SD outliers for removal, repeating this PCA process until all substantial outliers had been identified and removed. This PCA approach identified 37 sample outliers for removal.

As a final step in preparing the DS target dataset for PRS analysis, we removed the major histocompatibility complex region (Chr6: 25-34 Mb, hg19), which is a region of extended high linkage disequilibrium that can overly influence PRS results. Our final data set included 487 samples (245 DS+AVSD cases, 242 DS+NH controls) and 2,351,951 autosomal SNPs (excluding chromosome 21). The multiple steps involved in generating this final data set for the primary PRS analyses are presented as a flowchart in Figure 3.3.

Target dataset for secondary PRS analysis

Our secondary PRS analyses examined the contribution by alleles on the trisomic chromosome 21 to a polygenic component for DS-associated AVSD. To do this, we compared PRS results based on polygenic scores generated using all autosomes (including chromosome 21) to PRS results based on scores using all autosomes except for chromosome 21.

We analyzed the same set of target samples as for the primary analyses (245 DS+AVSD cases, 242 DS+NH controls), 158 of whom had WGS data for chromosome 21, and 329 of whom had Affymetrix Genome-Wide Human SNP 6.0 array genotype data for chromosome 21 (given the complexities of imputing trisomic genotypes, we did

not have imputed data for these 329 samples). Given that trisomic data cannot be represented by the PLINK1.9 binary format, we handled these chromosome 21 data separately from the other chromosomes. Prior to merging chromosome 21 data for these WGS and array samples, we applied certain QC filters. None of the 158 WGS samples nor the 329 array samples had an excess of missing genotypes for chromosome 21 (all had approximately 5% or less missingness). For variant QC, we excluded SNPs missing for > 5% of samples, as well as SNPs with A/T, T/A, C/G, and G/C alleles which can be difficult to match between datasets due to strand ambiguity. We also removed SNPs with significantly different allele frequencies between the WGS and array datasets (we determined that a frequency difference of ≥ 0.125 was an appropriate threshold for these chromosome 21 datasets). Post-merger, we removed SNPs with excess missingness specifically among cases or controls (missing for > 3% of cases or > 3% of controls), and we also excluded SNPs that were monoallelic in the full sample. These steps yielded a merged chromosome 21 dataset with 487 samples and 3,984 SNPs.

We then took the dataset used for the primary analyses (487 samples and 2,351,951 autosomal SNPs, excluding chromosome 21), and limited it to SNPs on the Affymetrix Genome-Wide Human SNP 6.0 array, leaving 389,544 SNPs. This was done since the chromosome 21 data were also necessarily limited to the array SNPs. We used these array-based genotype data, both with and without the chromosome 21 data, in order to perform the secondary PRS analyses.

Supplemental References

Anderson, C.A., F.H. Pettersson, G.M. Clarke, L.R. Cardon, A.P. Morris *et al.* 2010. Data quality control in genetic case-control association studies. *Nat Protoc* 5 (9):1564-1573.

Chang, C.C., C.C. Chow, L.C. Tellier, S. Vattikuti, S.M. Purcell *et al.* 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7.

Das, S., L. Forer, S. Schonherr, C. Sidore, A.E. Locke *et al.* 2016. Next-generation genotype imputation service and methods. *Nat Genet* 48 (10):1284-1287.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437 (7063):1299-1320.

Kent, W.J., C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle *et al.* 2002. The human genome browser at UCSC. *Genome Res* 12 (6):996-1006.

McCarthy, S., S. Das, W. Kretzschmar, O. Delaneau, A.R. Wood *et al.* 2016. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 48 (10):1279-1283.

Purcell, S., and C. Chang. PLINK [v1.9b6.6]. www.cog-genomics.org/plink/1.9/.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramachandran, D., Z. Zeng, A.E. Locke, J.G. Mulle, L.J. Bean *et al.* 2015. Genome-Wide Association Study of Down Syndrome-Associated Atrioventricular Septal Defects. *G3 (Bethesda)* 5 (10):1961-1971.

Rayner, W. Script to check plink .bim files against HRC/1000G for strand, id names, positions, alleles, ref/alt assignment [v 4.2.9]. https://www.well.ox.ac.uk/~wrayner/tools/.

Tables

Table 3.1. Summary of cohort for SKAT-O analysis

WES	**Cases**	**Controls**	**WGS**	**Cases**	**Controls**
Total	190 (174)	138 (126)	Total	169 (148)	39 (27)
Caucasian	152 (152)	101 (101)	Caucasian	161 (148)	35 (27)
African American	34 (18)	37 (25)	African American	7 (0)	4 (0)
Ad-mixed American	2 (2)	0	Ad-mixed American	0	0
East Asian	2 (2)	0	East Asian	0	0
Hispanic	0	0	Hispanic	1 (0)	0

Total cohort numbers after QC and PCA used for SKAT-O analyses in parentheses.

Table 3.2. Summary of genes analyzed using SKAT-O, based on variants in exons and UTR regions.

SKAT-O analysis	**MAF filter**	**Genes**	**SNVs**	**MAF weighting in SKAT**
WES - common	MAF > 0.05	10,228	25,355	no
WES - rare	MAF < 0.01 and missing in gnomAD	17,318	142,006	yes
WES - ultra-rare	MAF < 0.001	14,898	59,092	yes

Table 3.3: First discovery dataset: diagnoses for 2,594 mixed CHD cases (Cordell, Bentham, et al. 2013; Cordell, Töpf, et al. 2013)

CHD diagnosis	Number (%) of samples
Tetralogy of Fallot	835 (32.2)

Left-sided malformations	387 (14.9)
Ostium secundum atrial septal defect	340 (13.1)
Transposition of the great arteries	207 (8.0)
Ventricular septal defect	191 (7.4)
Conotruncal malformations	151 (5.8)
Double outlet right ventricle	96 (3.7)
AVSD (partial and complete)	73 (2.8)
Other CHD*	314 (12.1)

*For a more complete list of included CHD diagnosis, see (Cordell, Bentham, et al. 2013; Cordell, Töpf, et al. 2013)

Table 3.4: Second discovery dataset: diagnoses for 406 mixed CHD cases (Agopian et al., 2017)

CHD diagnosis	Number (%) of samples
Tetralogy of Fallot	134 (33.0)
Ventricular septal defect	109 (26.8)
D-transposition of the great arteries	80 (19.7)
Double outlet right ventricle	25 (6.2)
Isolated aortic arch anomalies	22 (5.4)
Truncus arteriosus	19 (4.7)
Other CHD	17 (4.2)

Table 3.5. SKAT-O results of common variants: Common variants are defined as MAF > 0.05 in gnomAD. All genes were tested in the WES dataset; only the top-ranked genes (p<0.001) were tested in the WGS dataset as a replication set. Common variant SKAT-O analyses were not weighted by MAF.

Gene	Loci	p-value - WES	Variants Tested	p-value - WGS	Variants Tested
AMOT	chrX:112,774,503-112,840,815	8.21E-04	1	0.849	6
CEP290	chr12:88,049,016-88,142,088	1.88E-04	3	0.064	3
CTDSP1	chr2:218,399,755-218,405,941	8.82E-04	2	0.392	3
DNAJA4	chr15:78,264,086-78,282,196	2.37E-04	2	0.296	7
GABRE	chrX:151,953,124-151,974,676	2.49E-04	1	0.339	6
HHAT	chr1:210,328,252-210,676,296	6.39E-04	3	0.576	5
HJURP	chr2:233,836,702-233,854,535	8.49E-04	7	0.592	10
MEFV	chr16:3,242,028-3,256,627	7.18E-04	8	0.388	12
MRGPRX3	chr11:18,120,955-18,138,480	6.07E-04	3	0.808	6
MYO5B	chr18:49,822,789-50,195,147	2.72E-04	5	0.263	14
NEK10	chr3:27,110,904-27,369,392	8.52E-04	3	0.092	8
NR0B2	chr1:26,911,489-26,913,975	5.91E-04	1	0.83	3
PLEKHM3	chr2:207,821,288-208,025,527	3.48E-04	2	0.826	10
SAG	chr2:233,307,816-233,347,055	3.35E-04	3	0.527	3
TRMT9B	chr8:12,945,673-13,029,777	6.22E-05	8	0.627	40
WDR61	chr15:78,283,235-78,299,609	5.98E-04	1	0.414	2
WDR87	chr19:37,884,932-37,906,677	8.81E-05	4	0.198	13
ZNF571	chr19:37,562,392-37,594,790	8.60E-04	3	0.203	5

ZNF573	chr19:37,738,302-37,779,590	7.08E-04	2	0.148	3

Table 3.6. SKAT-O results of rare variants: Rare variants are defined as MAF < 0.01 or missing in gnomAD with an additional dataset MAF filter < 0.02. All genes were tested in the WES dataset; only the top-ranked genes (p<0.001) were tested in the WGS dataset as a replication set.

Gene	**Loci**	**p-value - WES**	**Variants Tested**	**p-value - WGS**	**Variants Tested**
ALG11	chr13:52,012,398-52,033,600	9.66E-04	6	0.543	6
CST4	chr20:23,685,640-23,689,040	6.21E-04	8	0.815	5
NOTCH4	chr6:32,194,843-32,224,067	6.66E-04	9	0.031	10
PODNL1	chr19:13,933,957-13,953,302	7.86E-04	5	0.568	14
RBM12	chr20:35,648,925-35,664,900	9.04E-04	3	0.523	8
RNF135	chr17:30,971,039-30,999,911	9.14E-04	6	0.651	5
RNF152	chr18:61,808,067-61,893,007	5.06E-04	7	0.701	7
SLIT3	chr5:168,661,740-169,301,129	7.30E-04	23	0.484	23
TRIM56	chr7:101,085,481-101,097,967	4.73E-04	8	0.836	6
VCX3A	chrX:6,533,618-6,535,118	9.07E-04	4	0.69	4

Table 3.7. SKAT-O results of ultra-rare variants: Ultra-rare variants are defined as MAF < 0.001 in gnomAD without variants missing in gnomAD to test whether the well-defined ultra-rare variants are driving the top rare results. All genes were tested in the WES dataset; only the top-ranked genes (p<0.001) were tested in the WGS dataset as a replication set.

Gene	**Loci**	**p-value - WES**	**Variants Tested**	**p-value - WGS**	**Variants Tested**
ZNF318	chr6:43,336,070-43,369,647	8.07E-04	17	0.042	2

Table 3.8. SKAT-O results in WES dataset for the two pathways suggested by the single gene test results and by previous literature.

Gene	p-value - rare	Variants Tested	p-value - ultra-rare	Variants Tested	p-value - common	Variants Tested
Cilia pathway	0.04	3542	0.80	1490	0.77	674
Notch pathway	0.03	487	0.39	222	0.24	73

Table 3.9. PRS results using discovery GWAS of 2,594 mixed CHD cases and 5,159 controls and SNPs with MAF ≥ 0.35. 'Threshold' indicates that SNPs with discovery GWAS p-values below the threshold were used for PRS construction, and 'No. SNP' is the corresponding number of SNPs used for scoring. OR: Odds ratio per standard deviation increase in PRS, CI: Confidence interval, Nag. r_2: Nagelkerke's r_2, P_{unadj}: Uncorrected p-value, P_{adj}: P-value corrected for multiple correlated tests.

Threshold	No. SNP	OR	95% CI	Nag. r_2	P_{unadj}	P_{adj}
1e-05	1	1.12	0.91-1.38	0.24%	0.278	> 0.15
1e-04	5	1.19	0.96-1.47	0.54%	0.107	> 0.15
0.001	93	1.27	1.03-1.57	1.03%	0.027	> 0.15
0.005	328	1.25	1.01-1.54	0.91%	0.037	> 0.15
0.01	597	1.35	1.09-1.67	1.61%	0.006	> 0.15
0.05	2,421	1.25	1.02-1.54	0.95%	0.033	> 0.15
0.1	4,275	1.28	1.03-1.57	1.09%	0.023	> 0.15
0.2	7,590	1.22	0.99-1.50	0.75%	0.059	> 0.15
0.3	10,432	1.18	0.96-1.46	0.54%	0.108	> 0.15
0.4	12,982	1.11	0.91-1.37	0.22%	0.303	> 0.15
0.5	15,197	1.12	0.91-1.38	0.25%	0.278	> 0.15
1	22,507	1.09	0.89-1.34	0.15%	0.389	> 0.15

Figures

Figure 3.1. Representative SKAT-O Manhattan plot and QQ plot of common variants. Each dot represents a gene in the SKAT-O analysis, ordered by chromosome. No gene reached Bonferroni significance (red horizontal line), however 30 genes showed a nominal significance level of $p < 0.001$ (blue horizontal line).

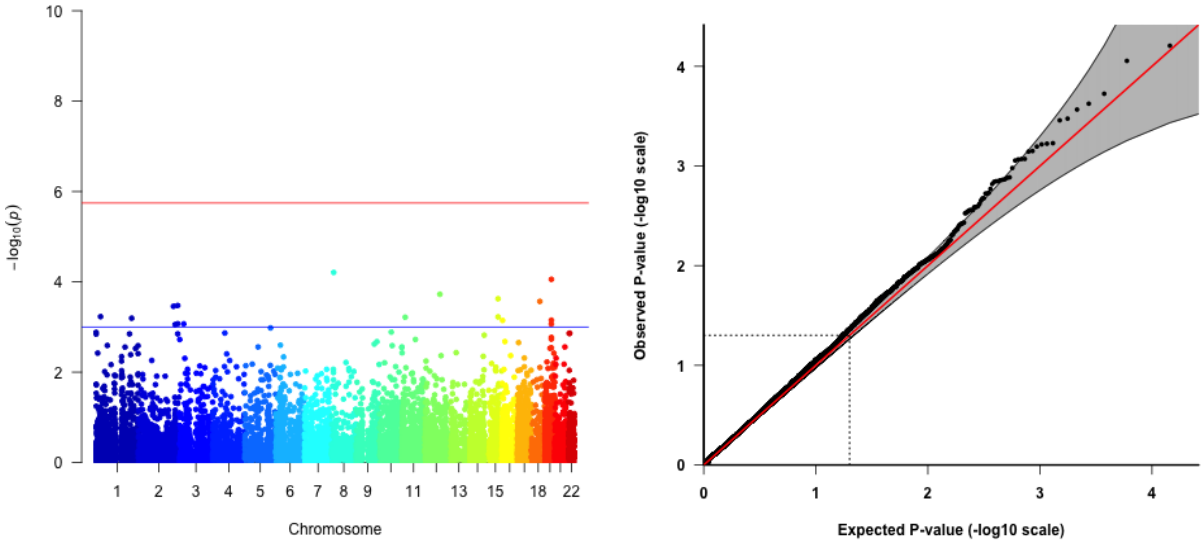


Figure 3.2. PRS results using discovery GWAS of 2,594 mixed CHD cases and 5,159 controls and SNPs with MAF ≥ 0.35. Plot shows odds ratio per standard deviation increase in PRS, with corresponding 95% confidence interval. 'P-value threshold' indicates that SNPs with discovery GWAS p-values below the threshold were used for PRS construction. P_{adj} is the p-value after correction for multiple correlated tests.

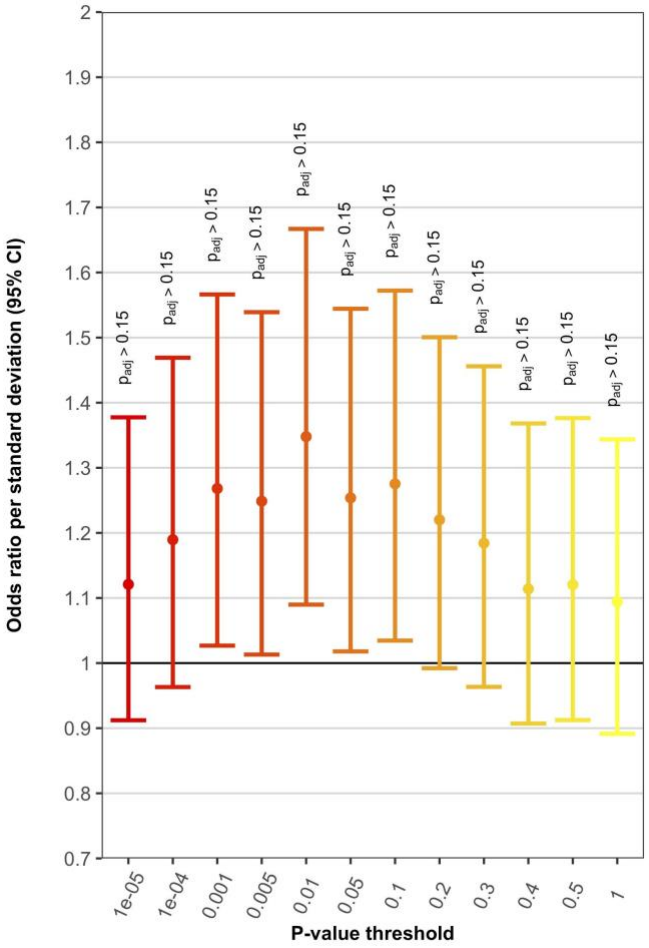


Figure 3.3. Flowchart showing the multiple steps involved in generating the final data set for the primary PRS analyses.

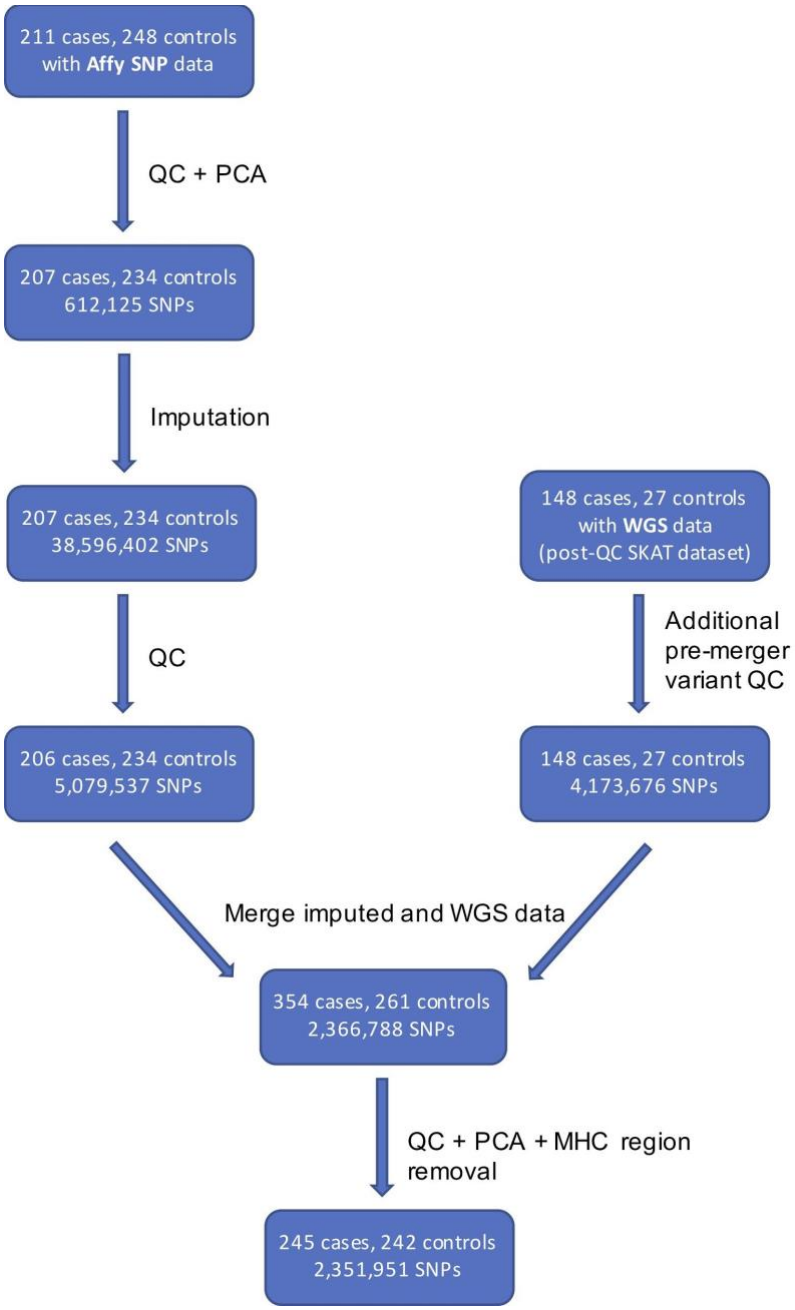

211 cases, 248 controls
with **Affy SNP** data
QC + PCA
207 cases, 234 controls
612,125 SNPs
Imputation
207 cases, 234 controls
38,596,402 SNPs
QC
206 cases, 234 controls
5,079,537 SNPs
148 cases, 27 controls
with **WGS** data
(post-QC SKAT dataset)
Additional
pre-merger
variant QC
148 cases, 27 controls
4,173,676 SNPs
Merge imputed and WGS data
354 cases, 261 controls
2,366,788 SNPs
QC + PCA + MHC region
removal
245 cases, 242 controls
2,351,951 SNPs

Figure 3.4. PRS results using discovery GWAS of 2,594 mixed CHD cases and 5,159 controls and various MAF thresholds. MAF thresholds were applied to the discovery GWAS; SNPs with MAF below the threshold were excluded from PRS construction. Top row: Each plot displays odds ratio per standard deviation in PRS and the corresponding 95% confidence interval (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). P_{adj} are adjusted p-values (corrected for multiple correlated tests). 95% CIs correspond to unadjusted p-values. Bottom row: Each plot displays Nagelkerke's r_2 (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). Numbers above each r_2 bar are the number of SNPs used to construct PRS at that particular p-value threshold and MAF filter combination.

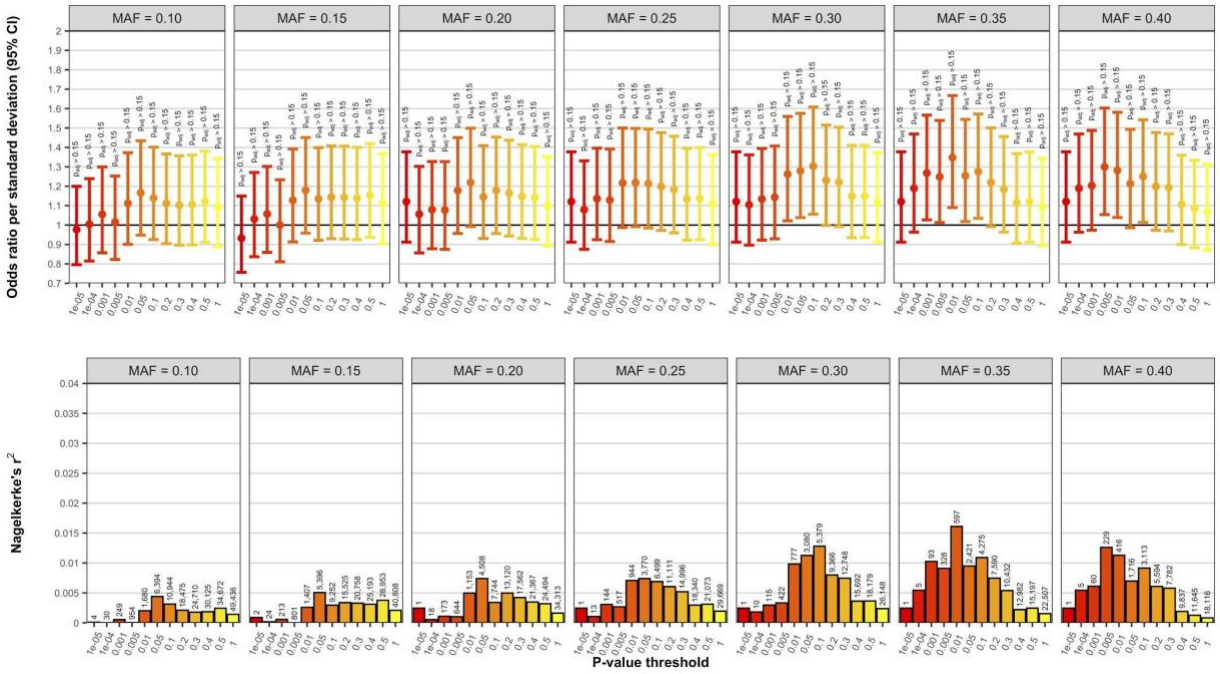


Figure 3.5. PRS results using discovery GWAS of 406 mixed CHD cases and 2,976 controls and various MAF thresholds. MAF thresholds were applied to the discovery GWAS; SNPs with MAF below the threshold were excluded from PRS construction. Top row: Each plot displays odds ratio per standard deviation in PRS and the corresponding 95% confidence interval (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). P_{adj} are adjusted p-values (corrected for multiple correlated tests). 95% CIs correspond to unadjusted p-values. Bottom row: Each plot displays Nagelkerke's r_2 (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). Numbers above each r_2 bar are the number of SNPs used to construct PRS at that particular p-value threshold and MAF filter combination.

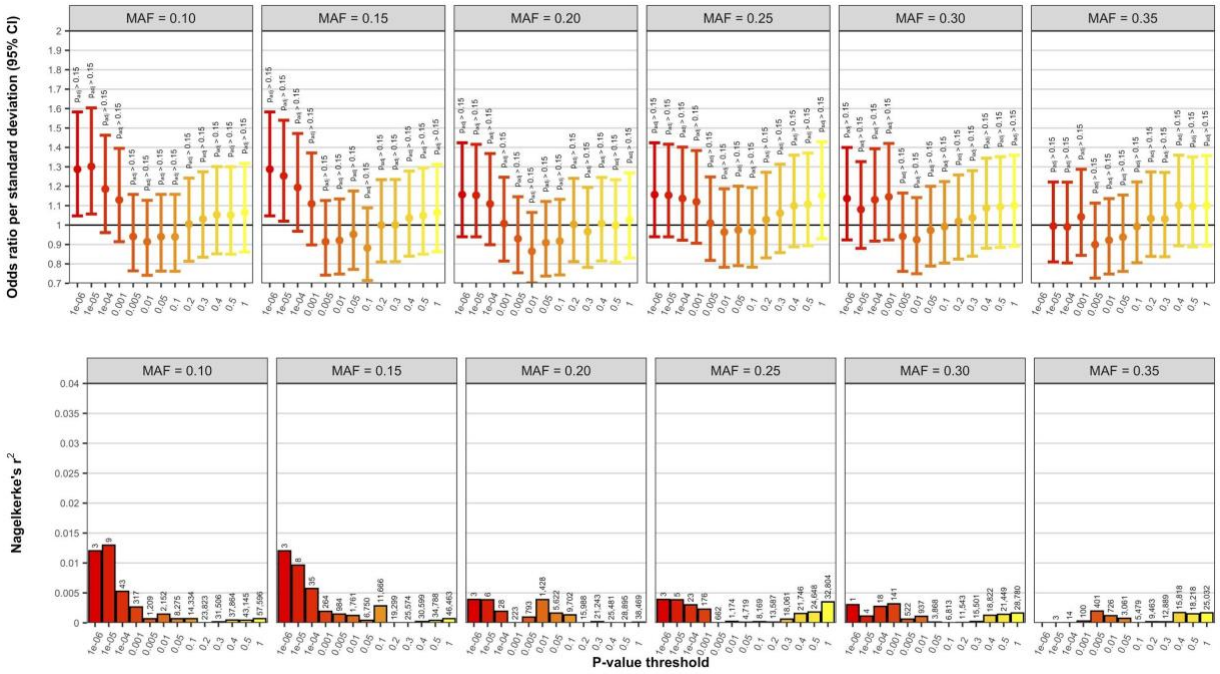


Figure 3.6. PRS results using meta-analysis of two GWAS as discovery dataset and employing inverse variance weighted SNP effects for scoring, for various MAF thresholds. MAF thresholds were applied to the discovery GWAS; SNPs with MAF below the threshold were excluded from PRS construction. Top row: Each plot displays odds ratio per standard deviation in PRS and the corresponding 95% confidence interval (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). P_{adj} are adjusted p-values (corrected for multiple correlated tests). 95% CIs correspond to unadjusted p-values. Bottom row: Each plot displays Nagelkerke's r_2 (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). Numbers above each r_2 bar are the number of SNPs used to construct PRS at that particular p-value threshold and MAF filter combination.

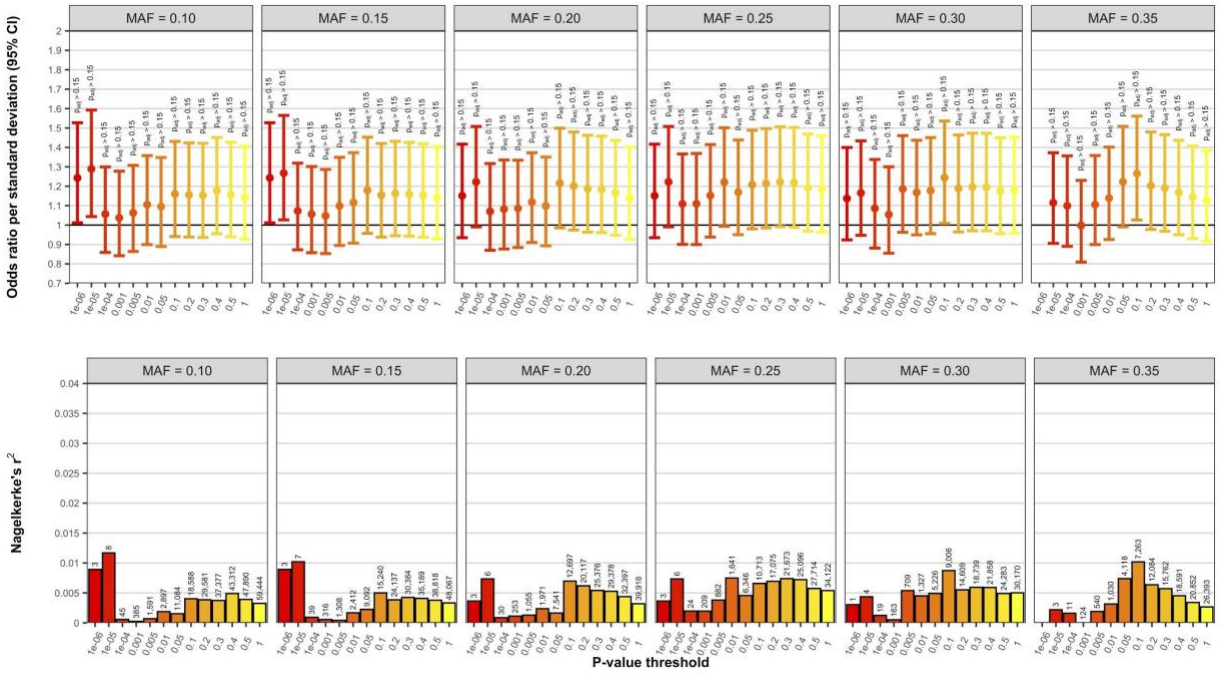


Figure 3.7. PRS results for all autosomes *excluding* chromosome 21. These analyses used the discovery GWAS of 2,594 mixed CHD cases and 5,159 controls. Various MAF thresholds were applied to the discovery GWAS; SNPs with MAF below the threshold were excluded from PRS construction. Top row: Each plot displays odds ratio per standard deviation in PRS and the corresponding 95% confidence interval (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). P_{adj} are adjusted p-values (corrected for multiple correlated tests). 95% CIs correspond to unadjusted p-values. Bottom row: Each plot displays Nagelkerke's r_2 (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). Numbers above each r_2 bar are the number of SNPs used to construct PRS at that particular p-value threshold and MAF filter combination.

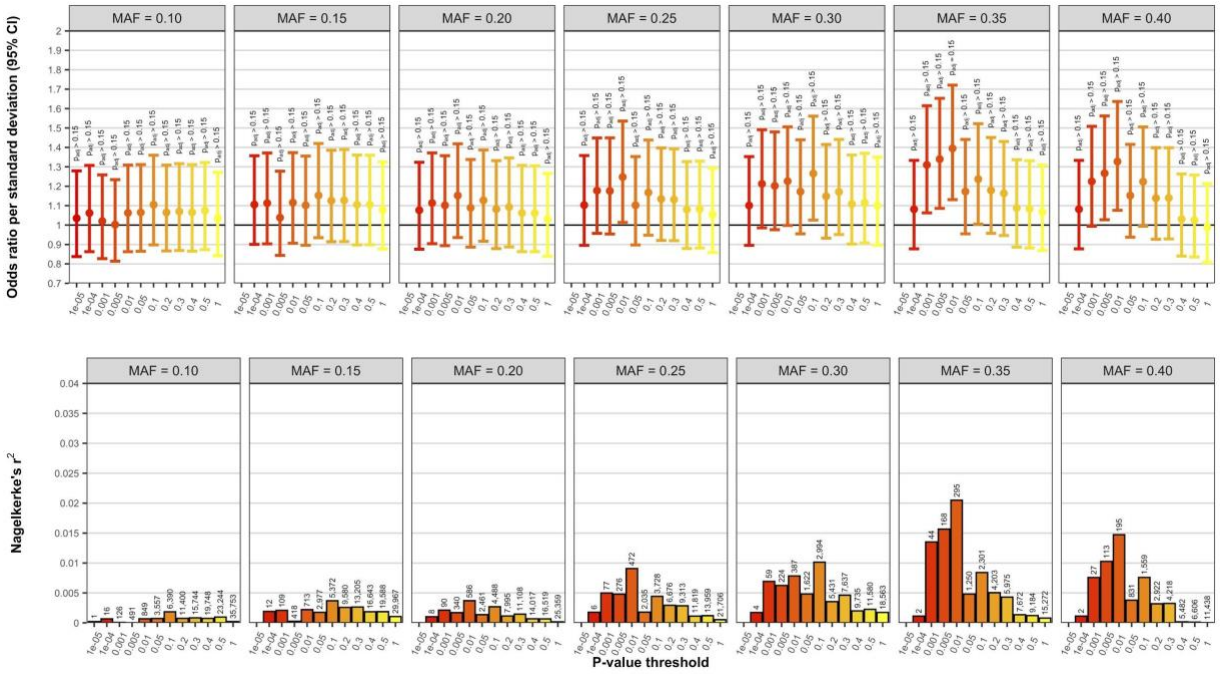


Figure 3.8. PRS results for all autosomes *including* chromosome 21. These analyses used the discovery GWAS of 2,594 mixed CHD cases and 5,159 controls. Various MAF thresholds were applied to the discovery GWAS; SNPs with MAF below the threshold were excluded from PRS construction. Top row: Each plot displays odds ratio per standard deviation in PRS and the corresponding 95% confidence interval (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). P_{adj} are adjusted p-values (corrected for multiple correlated tests). 95% CIs correspond to unadjusted p-values. Bottom row: Each plot displays Nagelkerke's r_2 (y-axis) for PRS constructed based on particular discovery GWAS p-value thresholds (x-axis). Numbers above each r_2 bar are the number of SNPs used to construct PRS at that particular p-value threshold and MAF filter combination.

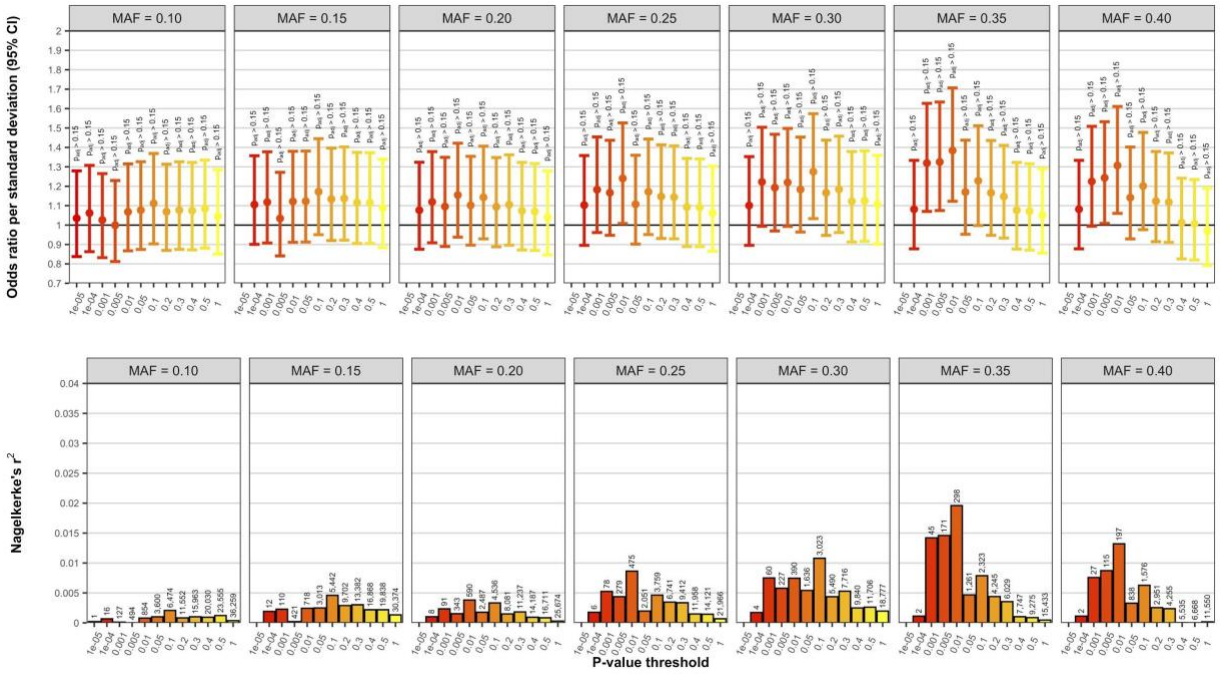
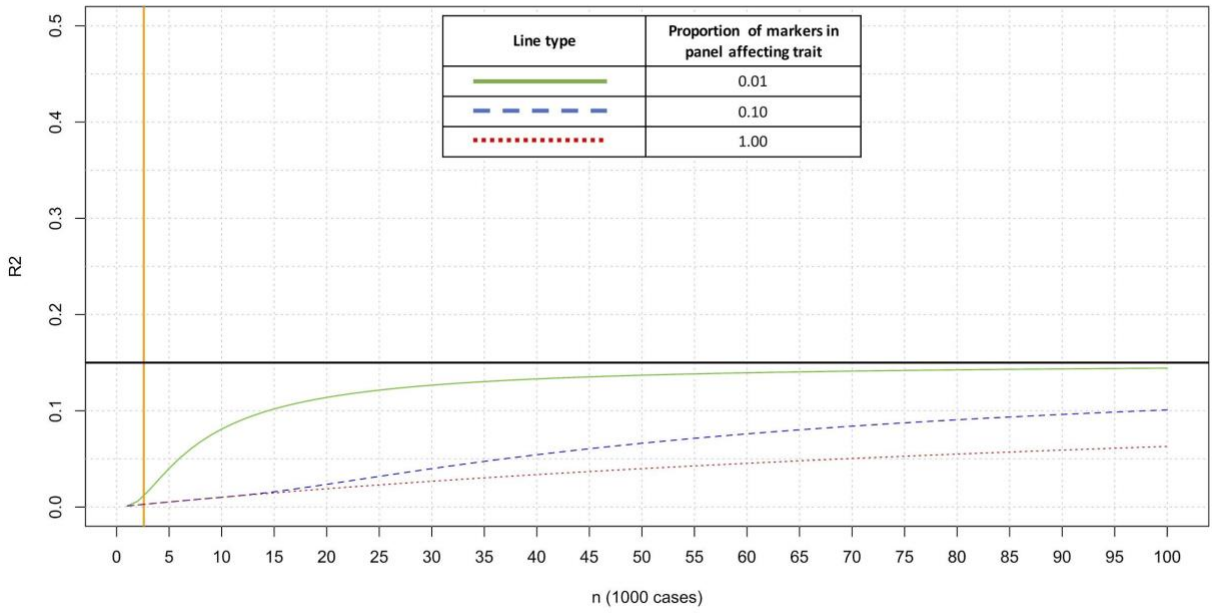
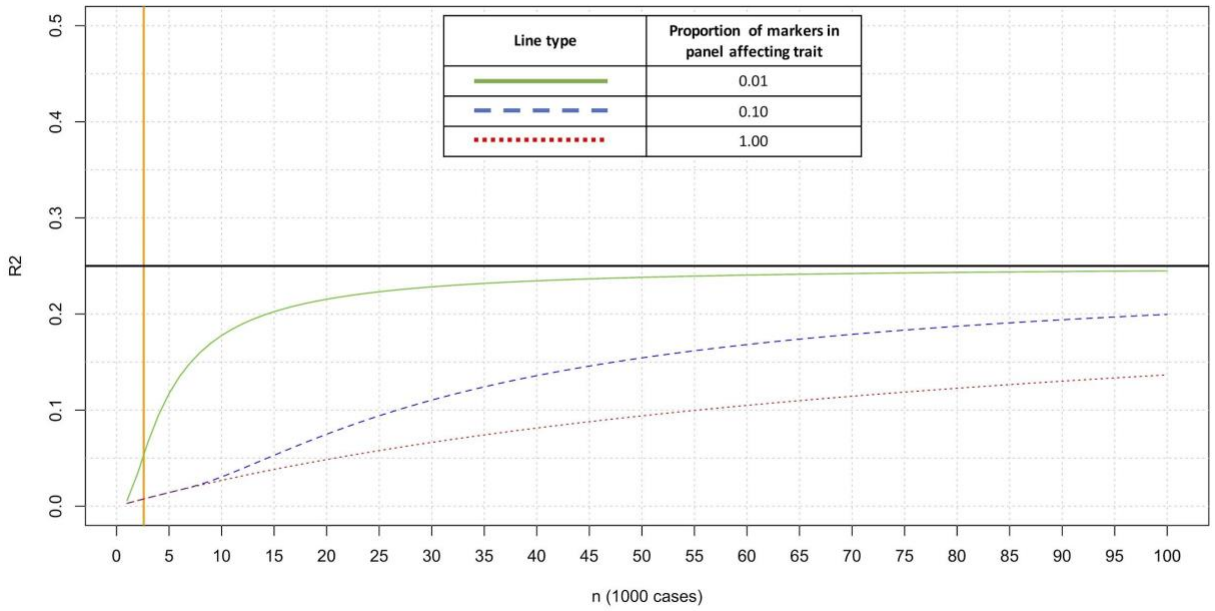


Figure 3.9 Maximum variance in target phenotype that can be explained by PRS (y-axis: liability scale r_2) given a range of training sample sizes (x-axis: number of cases in thousands). Assumptions: training sample with case:control ratio of 1:2 (same as ratio for larger of the two independent CHD discovery datasets); target sample with case:control ratio of 1:1 (same as ratio for DS target dataset); prevalence of CHD in training population is 1%; prevalence of AVSD in DS target population is 20%; 100,000 independent variants in the training SNP panel; genetic effects for training and target samples are identical (correlation = 1); proportion of SNPs in the training set panel that affect the training phenotype is 1%, 10% or 100%. For plot **A**, amount of variance in the training phenotype explained by the training set SNP panel (Vg_{train}) is 15%; for plot **B** Vg_{train} is 25%; for plot **C** Vg_{train} is 35%. Solid black horizontal line marks the maximum r_2 that can be explained by PRS using an infinitely large training sample size (given the assumed parameters). Vertical orange line marks the number of CHD cases in the larger of the two independent discovery datasets (2,594 cases).

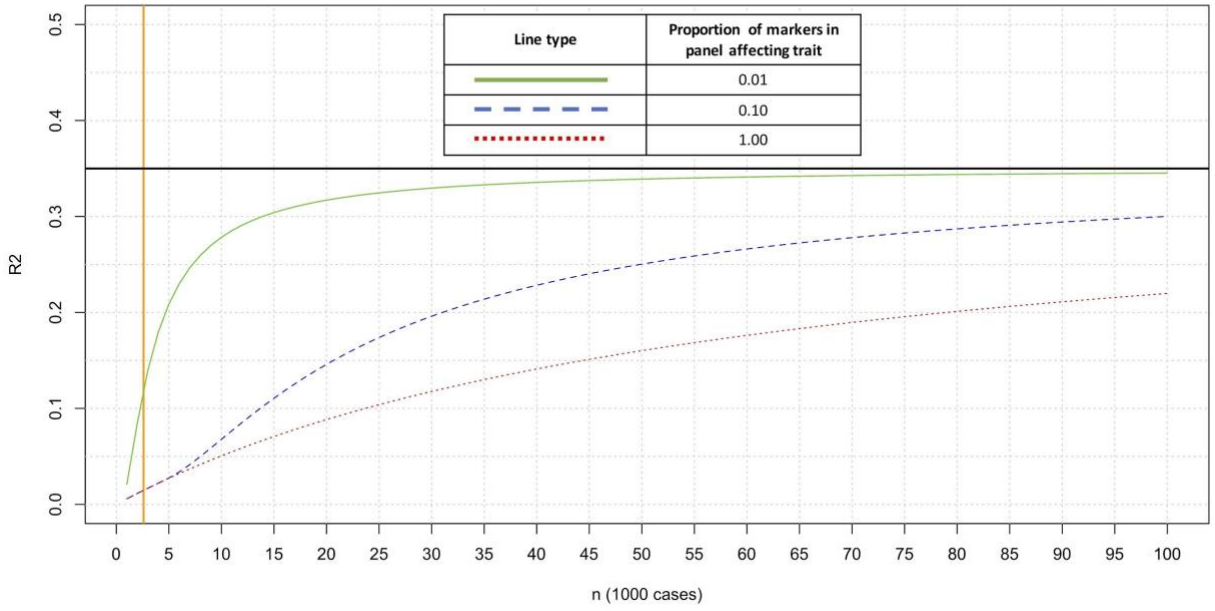
A.



B.


Line type
Proportion of markers in panel affecting trait
0.01
0.10
1.00
R2
0.0
0.1
0.2
0.3
0.4
0.5
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
n (1000 cases)

C.


Line type
Proportion of markers in panel affecting trait
0.01
0.10
1.00
R2
0.0
0.1
0.2
0.3
0.4
0.5
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
n (1000 cases)

IV. Bystro: rapid online variant annotation and natural-language filtering at whole-genome scale

Alex V. Kotlar, Cristina E. Trevino, Michael E. Zwick, David J. Cutler & Thomas S. Wingo

Previously published in:
Kotlar, A. V., et al. **Bystro: Rapid online variant annotation and natural-language filtering at whole-genome scale**. Genome Biology, 2017

Background

While genome-wide association studies (GWAS) and whole-exome sequencing (WES) remain important components of human disease research, the future lies in whole-genome sequencing (WGS), as it inarguably provides more complete data. The central challenge posed by WGS is one of scale. Genetic disease studies require thousands of samples to obtain adequate power, and the resulting WGS datasets are hundreds of gigabytes in size and contain tens of millions of variants. Manipulating data at this scale is difficult. To find the alleles that contribute to traits of interest, two steps must occur. First, the variants identified in a sequencing experiment need to be described in a process called annotation, and second, the relevant alleles need to be selected based on those descriptions in a procedure called variant filtering.

Annotating and filtering large numbers of variant alleles requires specialty software. Existing annotators, such as ANNOVAR (Chang & Wang, 2012), SeqAnt (Shetty et al., 2010), VEP (McLaren et al., 2016), and GEMINI (DeFreitas et al., 2016) have played an important research role, and are sufficient for small to medium experiments (e.g.,10s to 100s of WES samples). However, they require significant computer science training to use in offline, distributed computing environments, and have substantial restrictions in terms of performance and the maximum size of the data they will annotate online. Existing variant filtering solutions are even more limited, with most analyses requiring

researchers to program custom scripts, which can result in errors that impact reproducibility (Sandve et al., 2013). Therefore, annotation and filtering are not readily accessible to most scientists, and even bioinformaticians face challenges of performance, cost and complexity.

Here we introduce an application called Bystro that significantly simplifies variant annotation and filtering, while also improving performance by orders of magnitude and saving weeks of processing time on large data sets. It is the first program capable of handling sequencing experiments on the scale of thousands of whole-genome samples and tens of millions of variants online in a web browser, and integrates the first, to our knowledge, publicly-available, online natural-language search engine for filtering variants and samples from these experiments. The search engine enables real-time (sub-second), nuanced variant filtering, both across all samples and per sample, using simple phrases and interactive, web-based filters. Bystro makes it possible to efficiently find alleles of interest in any sequencing experiment without computer science training, improving reproducibility while reducing annotation and filtering costs.

Results

To compare Bystro's capabilities with other recent programs, we submitted 1000 Genomes (Auton et al., 2015) Phase 1 and Phase 3 VCF files for annotation and filtering (Figure 2.1). Phase 1 contains 39.4 million variants from 1,092 WGS samples, while Phase 3 includes 84.9 million alleles from 2,504 WGS samples. We first evaluated the online capabilities of the web-based versions of Bystro, wANNOVAR (Chang & Wang, 2012), VEP, and GEMINI (running on the Galaxy (Goecks et al., 2010) platform). Bystro was the only program able to complete either 1000 Genomes Phase 1 or Phase

3 online, and was also the only application to handle a $6x10^6$ variant subset of Phase 3, a size representative of modest whole-genome experiments. When tested with $5x10^4$ – $1x10^6$ variant subsets of 1000 Genomes Phase 3, Bystro was approximately 144 – 212x faster than GEMINI/Galaxy in generating a downloadable annotation and searchable result database, and was significantly easier to use, as it did not require a separate annotation step (Figure 2.2). When tested on a small trio data set, Bystro was able to identify *de novo* variants without any additional software, and was 45x faster than GEMINI's de_novo tool (Additional file 2.1: Table S2.1). Bystro and GEMINI/Galaxy produced similarly detailed outputs, with Bystro offering fewer, but more complete and recent sources, as well as more detailed annotations for some classes of data (Additional file 2.1: Table S2.2 ; Additional file 2.2). Notably GEMINI was found to work only with the hg19 human genome assembly, whereas Bystro supports hg19, hg38, and a variety of model organisms.

We next tested offline performance on identical servers to gauge performance in the absence of web-related file-size and networking limitations. Bystro was 113x faster than ANNOVAR and up to 790x faster than VEP, annotating all $8.5x10^7$ variants and 2,504 samples from Phase 3 in less than 3 hours (Table 2.1). Furthermore, ANNOVAR was unable to finish either Phase 1 or Phase 3 annotations due to memory requirements (exceeding 60GB of RAM), and VEP annotated Phase 3 at a rate of 10 variants per second, indicating that it would need at least 98 days to complete. Critically, Bystro's run time grew linearly with the number of submitted genotypes, suggesting that it could handle even hundreds of thousands of samples within days.

While offering significantly faster performance, Bystro also provided 3.5x the number of annotation output fields as ANNOVAR and 5.6x that of VEP (Additional file 2.3**).** Notably, unlike ANNOVAR or VEP, Bystro annotated each sample relative to its genotype, reporting homozygosity, heterozygosity, missingness, sample minor allele frequency, and labeling each sample as homozygous, heterozygous, or missing. In contrast, ANNOVAR provided only sample minor allele frequency, while VEP reported no sample-level data. We note that VEP is capable of providing per-sample annotations (heterozygosity/homozygosity status), but we were unable to use this feature for performance reasons. A detailed comparison of the exact settings used is given (Additional file 2.2 ; Additional file 2.3).

To investigate annotation accuracy, we next compared Bystro with ANNOVAR and VEP on a previously-analyzed synthetic dataset (Yen et al., 2017). Overall, excellent concordance between all methods was noted (Additional files 2.4, 2.5, and 2.6). For instance, in comparison with ANNOVAR, allele position (>98%), allele identity (100%), and variant effects (>99%) were highly consistent across all classes of variation, for sites that Bystro did not exclude for quality reasons (Additional file 2.4).

In cases where the annotators disagreed, Bystro gave the correct interpretations. For instance, Bystro and VEP excluded reference sites (ALT: “.”), while ANNOVAR annotated such loci as “synonymous SNV”; it is of course incorrect to call reference sites variant (Additional file 2.4 ; Additional file 2.5). In cases of insertions and deletions, which are often ambiguously represented in VCF files due to the format’s padding requirements, Bystro always provided the parsimonious left-shifted representation, while ANNOVAR and VEP occasionally right-shifted variants (Additional file 2.4 ; Additional

file 2.5). This is evident at chr15:42680000CA>CAA, where both ANNOVAR and VEP called the insertion as occuring after the first “A”, with 2 bases of padding, rather than the simpler option after the first base, “C”, with 1 base of padding (Additional file 2.1: Table S2.3). Similar results were found at multiallelic loci with complex indels (Additional file 2.1: Table S2.4).

Similarly, in cases where Bystro and ANNOVAR or VEP disagreed on variant consequences, Bystro always appeared correct relative to the underlying transcript set. For example, in the case of the simple insertion chr19:41123094G>GG, Bystro correctly identified all three overlapping transcripts (NM_003573;NM_001042544;NM_001042545), and noted the variant as coding (exonic) relative to all three. In contrast, ANNOVAR called the allele as disrupting a splice site, despite the fact that the nerest intron, and therefore splice site, was 37bp downstream (Additional file 2.1: Figure S2.1).

Additionally, Bystro’s strict VCF quality control measures substantially improved annotation accuracy.This is evident in the case of gnomAD, a VCF-format dataset that represents the largest experiment on human genetic variation. While Bystro and ANNOVAR provided identical gnomAD data for 93.7% of tested alleles, the remaining 6.3% were low-quality gnomAD results that were included in ANNOVAR and excluded from Bystro (Additional file 2.4). For instance, in the case of chr16:2103394C>T, ANNOVAR reported rs760688660, which failed gnomAD’s random forest quality control (QC) step. We note that a 6.3% false-positive rate is similar to the frequency of common variation, and significantly larger than the frequency of rare variants, making

ANNOVAR's gnomAD annotations a potentially unreliable source of data for both common and rare variant filtering.

Next, we explored the Bystro search engine's ability to filter the 84.9 million annotated Phase 3 variants. Bystro's search engine was unique in its natural-language capabilities, and no other tested online program could handle the full Phase 3 dataset, or subsets as large as $6x10_6$ variants (Figure 2.2). First, we used Bystro's search engine to find all alleles in exonic regions by entering the term "exonic" (933,343 alleles, 0.030 ± .001 seconds, Table 2.2). The search engine calculated a transition to transversion ratio of 2.96 for the query, consistent with previously observed values in coding regions**.** To refine results to rare, predicted deleterious alleles, we queried "cadd > 20 maf < .001 pathogenic expert review missense" (65 alleles, 0.029 ± 0.025s, Table 2.2). This search query could be written using partial words ("pathogen"), possessive nouns ("expert's"), different tenses ("reviews"), and synonyms ("nonsynonymous") without changing the results.

To test the search engine's ability to accurately match variants from full-text disease queries, we first searched "early-onset breast cancer", returning the expected alleles in *BRCA1* and *BRCA2* (4,335 variants, .037 ± .020s, Table 2)*.* Notably, the queried phrase "early-onset breast cancer" did not exist within the annotation, and instead matched closely-related RefSeq transcript names, such as "Homo sapiens breast cancer 2, early onset (BRCA2), mRNA**.**" We next explored Bystro's ability to handle synonyms and acronyms. To test the hypothesis that Bystro could interpret common ontologies, we queried "pathogenic nonsense E.D.S", where "nonsense" is a common synonym for "stopGain" (a term annotated by the Bystro annotation engine), and "E.D.S" is an

acronym for “Ehlers-Danlos Syndrome”. Bystro successfully parsed this query, returning a single *PLOD1* variant found in 1000 Genomes Phase 3 that introduces an early stop codon in all three of its overlapping transcripts, and which has been reported in Clinvar as “pathogenic” for “Ehlers-Danlos syndrome, type 4” (1 variant, .038s ± .027s, Table 2.2).

Since no other tested program could load or filter the 1000 Genomes Phase 3 VCF file online, we next compared Bystro to GEMINI (running on the Galaxy platform) on subsets of 1000 Genomes Phase 3. In contrast with GEMINI’s structured SQL queries, Bystro enabled shorter and more flexible searches. For instance, to return all missense, rare variants with CADD Phred scores larger than 15, GEMINI required a 162 character SQL query, while Bystro needed only 36 characters. Bystro also demonstrated synonym support, returning identical results for “missense” and “nonsynonymous” queries. Critically, Bystro’s search engine enabled real-time (sub-second) filtering, performing approximately four orders of magnitude faster than GEMINI on Galaxy while searching and returning similar volumes of data (Table 2.3).

To test the accuracy of Bystro’s search engine relative to the underlying annotation, we first compared Bystro’s natural-language queries with Bystro’s “Filters”, which provide a complimentary, exact-match filtering option. All results were identical between the two methods (Additional file 2.1: Table S2.5). To control for the possibility that Bystro’s “Filters” were biased, we created separate Perl filtering scripts that searched for exact matches within the underlying tab-delimited text annotation. Again, results were completely concordant (Additional file 2.1: Table S5). Finally, to control for the possibility that both Bystro’s “Filters” and the Perl scripts were biased due to the programmer, we

compared Bystro's natural-language queries with Excel filters on a smaller dataset that could be manually examined. The queries were found completely specific in this comparison as well (Additional file 2.1: Table S2.6; Additional file 2.7).

Discussion

The Bystro annotation and filtering capabilities are primarily exposed through a public web application (https://bystro.io/), and are also available for custom, offline installation. To ensure data safety, Bystro follows industry recommendations for password management, in-transit data security, and at-rest data security. Input and output files are encrypted at rest on Amazon EFS file systems, using AES 256-bit encryption, and every request for annotation or search data is authenticated by the web server using short-lived identity tokens. To further protect user data, annotation and search services are not directly open to the Internet, but require routing and authentication through the web server. Furthermore, all web traffic is encrypted using TLS (HTTPS), and password hashing follows the National Institute of Standards and Technology (NIST) recommended PBKDF2-HMAC-SHA512 strategy.

Creating an annotation online is as simple as selecting the genome and assembly used to make the variant call format (VCF) (Danecek et al., 2011) or SNP (Johnston et al., 2017) format files, and uploading these files from a computer or Amazon S3 bucket, which can be easily linked to the web application. Annotation occurs in the cloud, where distributed instances of the Bystro annotation engine process the data and send the results back to the web application for storage and display (Figure 2.1).

The Bystro annotation engine is open source, and supports diverse model organisms including *Homo sapiens* (hg19, hg38), *M. musculus* (mm9, mm10), *R. macaque*

(rheMac8), *R. norvegicus* (rn6), *D. melanogaster* (dm6), *C. elegans* (ce11), *S. cerevisiae* (sacCer3). To annotate, it rapidly matches alleles from users' submitted files to descriptions from RefSeq (O'Leary et al., 2016), dbSNP (Sherry et al., 2001), PhyloP (Pollard et al., 2010), PhastCons (Pollard et al., 2010), Combined Annotation-Dependent Depletion (CADD), Clinvar (Landrum et al., 2016), and gnomAD (Lek et al., 2016). For custom installations, Bystro supports Ensembl, RefSeq, or UCSC Known Genes transcript sets, and can be flexibly configured include annotations from any files in genePredExt, wigFix, BED, or VCF formats.

The annotation engine is aware of alternate splicing, and annotates all variants relative to each alternate transcript. When provided sample information, Bystro also annotates all variants relative to all sample genotypes. In such cases, at every site it labels each sample as homozygous, heterozygous, or missing, and also calculates the heterozygosity, homozogosity, missingness, and sample minor allele frequency. Furthermore, in contrast with current programs that require substantial VCF file pre-processing, Bystro automatically removes low-quality sites, normalizes variant representations, splits multi-allelic variants, and checks the reference allele against the genome assembly. Critically, Bystro's algorithm guarantees parsimonious (left-shifted) variant representations, even for multi-allelic sites containing complex insertions and deletions.

The Bystro annotation engine is designed to scale to any size experiment, offering the speed of distributed computing solutions such as Hail (Ganna et al., 2016), but with less complexity. Current well-performing annotators - such as ANNOVAR and SeqAnt - load significant amounts of data into memory to improve performance. However, when these

programs use multiple threads to take advantage of multicore CPUs they may exceed available memory (in some cases over 60GB), resulting in a sharp drop in performance or system crash. To solve this, Bystro annotates directly from an efficient memory-mapped database (LMDB), using only a few megabytes per thread, and because memory-mapped databases naturally lend themselves to the caching frequently accessed data, Bystro achieves most of the benefits of in-memory solutions, but without the per-thread penalties. This approach allows Bystro to take excellent advantage of multicore CPUs, while also enabling it to perform well on inexpensive, low-memory machines. Critically, when multiple files are submitted to it simultaneously, the Bystro annotation engine can automatically distribute the work throughout the cloud (or a user-configured computer cluster), gaining additional performance by processing the files on multiple computers (Figure 4.1). Furthermore, in reflection of the large sizes of both input sequencing experiments and the corresponding annotation outputs - on the order of terabytes for modern whole-genome experiments - Bystro accepts compressed input files, and directly writes compressed outputs. This ability to directly write compressed annotations with no uncompressed intermediate is critical given the rapid growth in sequencing experiment size.

When the web application receives a completed annotation, it saves the data and creates a permanent results page. Detailed information about the annotation, such as the database version used for the annotation is stored in a log file that the user may download. Users may then explore several quality control metrics, including the transition to transversion ratio on a per-sample or per-experiment basis. They may also download the results as tab-delimited text to their computer, or upload them to any

connected Amazon S3 bucket. In parallel with the completion of an annotation, the Bystro search engine automatically begins indexing the results. Once finished, a search bar is revealed in the results page, allowing users to filter their variants using the search engine (Figure 4.1).

Unlike existing filtering solutions, Bystro's Elasticsearch-based natural-language search engine accepts unstructured, "full-text" queries, and relies on a sophisticated language parser to match annotated variants. This allows it to offer the flexibility of modern search engines like Google and Bing, while remaining specific enough for the precise identification of alleles relevant to the research question. The Bystro search engine matches terms regardless of capitalization, punctuation, or word tense, and accurately finds partial terms within long annotation values. Like the annotation engine, the search engine is also exceptionally fast, automatically distributing indexed annotations throughout the cloud, enabling users to sift through millions of variants from large whole-genome sequencing experiments in milliseconds.

In order to provide flexible, but specific matches without relying on structured SQL queries, the search engine identifies the data type of every value in the annotation. Text undergoes stemming and lemmatization, which reduces the influence of grammatical variation, and is then tokenized into left-edge n-grams, which allows for flexible matching. Numerical data is stored in the smallest integer or float format that can accommodate it, allowing for rapid and accurate range queries. For complex queries, the search engine supports Boolean operators (AND, OR), regular expressions, and Levenshtein-edit distance fuzzy matches. It also has a built-in dictionary of synonyms, for instance equating "stopgain" and "nonsense".

In some cases, text will match accurately, but not specifically; this most often happens with short, generic terms. For instance, querying “intergenic” alone may match the word “intergenic” in “long intergenic non-protein coding RNA” in refSeq’s description field**,** as well as “intergenic” in the refSeq’s siteType field. To help improve accuracy in such cases, Bystro provides three, closely related features: 1) “Aggregations” allows users to see the top 200 values for any text field, or equivalently the min, max, mean, standard deviation (and other similar statistics) for any numerical field. This allows users to quickly and precisely understand the composition of search results, as well as to generate summary statistics. 2) “Filters” allows users to refine queries, by forcing the inclusion or exclusion of any values found in any field. For instance, rather than query “intergenic”, it may be easier and more precise to simply click on the “refSeq.siteType” filter, and select the “intergenic” value. Any number of “Filters” may be combined with any natural-language query, containing up to 1 million words. 3) Bystro allows field names within a natural-language query for added specificity. For example, rather than searching for “intergenic”, the user could type “refSeq.siteType:intergenic”, to indicate that they wished to match “intergenic” specifically in the refSeq.siteType annotation field.

Bystro’s search engine also includes several features to increase flexibility beyond the contents of the annotation: 1) “Custom Synonyms” allows users to define their own terms and annotations. Among other uses, this make it is possible to label trios, which can be used to easily identify *de novo* variants and test allele transmission models. 2) “Search Tools” are small programs, accessible by a single mouse click, that dynamically modify any query to generate complex result summaries. Some of their functions

include identifying compound heterozygotes. 3) “Statistical Filters” dynamically perform statistical tests on the variants returned from any query. For instance, the “HWE” filter allows users to exclude variants out of Hardy-Weinberg Equilibrium. This is an often-needed quality control step.

Most importantly, there is no limit to the number of query terms and “Filters” that can be combined, and users can save and download the results of any search query, which enables recursive filtering on a single dataset. The saved results are indexed for search, and hyperlinked to the annotations that they were generated from, forming permanent records that can be used to reproduce complex analyses. This multi-step filtering provides functionality similar to custom command-line filtering script pipelines, but is significantly faster, less error prone, and accessible to researchers without programming experience.

While Bystro’s annotation and filtering performance is currently unparalleled by any other approach, other software (such as Hail (Ganna et al., 2016)) could achieve similar performance by implementing distributed computing algorithms like MapReduce (Taylor, 2010), and spreading annotation workloads across many servers. Bystro demonstrates that these workarounds are unnecessary to achieve reasonable run-times for large datasets online or offline. Additionally, while Bystro’s natural-language search engine significantly reduces the difficulty of variant filtering, it does not handle language idiosyncrasies as robustly as more mature solutions like Google’s, and may return unexpected results when search queries are very short and non-specific, since such queries may have multiple correct matches. This is easily avoided by using longer phrases, by using “Custom Synonyms” to define more specific terms, by examining the

composition of results using “Aggregations”, or by applying “Filters” to precisely filter results. Such considerations and options are well-documented in Bystro’s online user guide (https://bystrio.io/help).

Conclusions

To date, identifying alleles of interest in sequencing experiments has been time-consuming and technically challenging, especially for whole-genome sequencing experiments. Bystro increases performance by orders of magnitude and improves ease of use through three key innovations: 1) a low-memory, high-performance, multithreaded variant annotator that automatically distributes work in cloud or clustered environments; 2) an online architecture that handles significantly larger sequencing experiments than previous solutions; and 3) the first publicly-available, general-purpose, natural-language search engine for variant filtering in individual research experiments. Bystro annotates large experiments in minutes, and its search engine is capable of matching variants within whole-genome datasets in milliseconds, enabling real-time data analysis. Bystro’s features enable practically any researcher – regardless of their computational experience - to analyze large sequencing experiments (e.g. thousands of whole-genome samples) within less than a day, and small ones (e.g. hundreds of whole-exome samples) in seconds. As genome sequencing continues the march toward ever-larger datasets and becomes more frequently used in diverse research settings, Bystro’s combination of performance and ease of use will prove invaluable for reproducible, rapid research.

Methods

Accessing Bystro

For most users, we recommend the Bystro web application (https://bystro.io), as it gives full functionality, supports arbitrarily large datasets, and provides a convenient interface to the natural-language search engine. Users with computational experience can download the Bystro open-source package (https://github.com/akotlar/bystro). Using the provided installation script or Amazon AMI image, Bystro can be easily deployed on an individual computer, computational cluster, or any Amazon Web Services (AWS) EC2 instance. Bystro has very low memory and CPU requirements, but benefits from fast SSD drives. As such we recommend at AWS instances with provisioned I/O EBS drives, RAID 0 non-provisioned EBS, or i2/i3-class EC2 instances.

Detailed documentation on Bystro's use, as well as example search queries can be found at https://bystro.io/help.

Bystro Database

Bystro databases were created using the open-source package (https://github.com/akotlar/bystro). The hg19 and hg38 databases contains RefSeq, dbSNP, PhyloP, PhastCons, Combined Annotation-Dependent Depletion (CADD), and Clinvar fields, as well as custom annotations (Additional file 2.8). A complete listing of the original source data is enumerated in the Git repository (https://github.com/akotlar/bystro/tree/master/config). Other organism databases contain

a subset of these sources, based on availability. Pre-built, up-to-date versions of these databases are publicly available (https://github.com/akotlar/bystro).

WGS Datasets

Phase 1 and Phase 3 autosome and chromosome X VCF files were downloaded from http://www.internationalgenome.org/data/. Phase 1 files were concatenated using bcftools (Li, 2011) "concat" function. Phase 3 files were concatenated using a custom Perl script (https://github.com/wingolab-org/GenPro/blob/master/bin/mergeSnpFiles). The Phase 1 VCF file was 895GB (139GB compressed), and the Phase 3 data was 853GB (15.6GB compressed). The larger size of Phase 1 can be attributed to the inclusion of extra genotype information (the genotype likelihood). The full Phase 3 chromosome 1 VCF file ($6.4x10_6$ variants, 1.2GB compressed), and $5x10_4$-$4x10_6$ variant allele subsets (8-655MB compressed) were also tested. All Phase 1 and Phase 3 data correspond to the GRCh37/hg19 human genome assembly. All data used are available (Additional file 2.9).

Online annotation comparisons

For online comparisons, the latest online versions offered at time of writing were used. Bystro beta10 (September 2017), wANNOVAR (April 2017), VEP (April 2017), and GEMINI (Galaxy version 0.8.1, released February 2016, latest as of October 2017) were tested online with the full 1000 Genomes Phase 1 and Phase 3 VCF files, unless they were unable to upload the files due to file size restrictions (Additional file 2.2). Bystro

was found to be the only program capable of uploading and processing the full Phase 1 and Phase 3 data sets, or subsets of Phase 3 larger than $1x10^6$ variants.

To conduct Bystro online annotations, a new user was registered within the public Bystro web application (https://bystro.io/). Phase 1 and Phase 3 files were submitted in triplicate, one replicate at a time, using the default database configuration (Additional file 2.2). Indexing was automatically performed by Bystro upon completion of each annotation. The Phase 3 annotation is publicly available to be tested (https://bistro.io/public).

The public Bystro server was configured on an Amazon i3.2xlarge EC2 instance. The server supported 8 simultaneous users. Throughout the duration of each experiment, multiple users had concurrent access to this server, increasing experiment variance, and limiting observed performance.

Online Variant Effect Predictor (VEP) submissions were done using the VEP web application (http://www.ensembl.org/info/docs/tools/vep/index.html). VEP has a 50MB (compressed) file size limit. Due to gateway timeout issues and this file size limit, data sets larger than $5x10^4$ variants failed to complete (Additional file 2.2).

Online ANNOVAR submissions were handled using the wANNOVAR web application. wANNOVAR could not accept the smallest tested file, the $5x10^4$ variant subset of Phase 3 chromosome 1 (8MB compressed) due to file size restrictions (Additional file 2.2).

Galaxy submission was made using the public Galaxy servers. Galaxy provides ANNOVAR, but its version of this software failed to complete any annotations, with the error “unknown option: vcfinput”. Annotations on Galaxy were therefore performed using GEMINI, which provides annotations similar to Bystro’s. Galaxy has a total storage

allocation of 250GB (after requisite decompression), and both Phase 1 and Phase 3 exceed this size. Galaxy was therefore tested with the full $6.4x10^6$ variant Phase 3 chromosome 1 VCF file. Galaxy's FTP server was able to upload the file; however, Galaxy was unable to load the data into GEMINI, terminating after running for 36 hours, with the message "This job was terminated because it ran longer than the maximum allowed job run time" (Additional file 2.2). Subsets of Phase 3 chromosome 1 containing $5x10^4$, $3x10^5$, and $1x10^6$ variants were therefore tested. Three repetitions of the $5x10^4$ variant submission were made. In consideration of the duration of execution, two repetitions were made of the $3x10^5$ and $1x10^6$ variants submissions. Since Galaxy does not record completion time, QuickTime was used to record each submission.
Bystro, VEP, and GEMINI online annotation times included the time to generate both a user-readable tab-delimited text annotation and a searchable database. GEMINI required an extra step to do so, using the query SELECT * FROM variants JOIN variant_impacts ON variants.name = variant_impacts.name.

Variant filtering comparisons

After Bystro completed each annotation, it automatically indexed the results for search. The time taken to index this data was recorded. Once this was completed, the Bystro web application's search bar was used to filter the annotated sequencing experiments. The query time, as well as the number of results and the transition to transversion ratio for each query, were automatically generated by the search engine and recorded. Query time did not take into account network latency between the search server and the

web server. All queries were run six times and averaged. The public search engine, which processed all queries, was hosted on a single Amazon i3.2xlarge EC2 instance. Since VEP, wANNOVAR, and Galaxy/GEMINI could not complete Phase 1 or Phase 3 annotations, variant filtering on these data sets could not be attempted. For small experiments VEP and GEMINI can filter based on exact matches, while wANNOVAR provides only pre-configured phenotype and disease model filters. VEP could annotate and filter at most only $5x10^4$ variants and was therefore excluded from query comparisons.

Galaxy/GEMINI was tested with subsets of 1000 Genomes Phase 3 of $1x10^6$ variants (the largest tested data set that Galaxy could handle), with the described settings (Additional file 2.2). In all GEMINI queries a JOIN operation on the variant_impacts table was used to return all variant consequences, and all affected transcripts, as Bystro does by default. Similarly, Bystro's CADD query was restricted to single nucleotide polymorphisms (using alt:(A || C || T || G)), as its behavior diverges from GEMINI's at insertions and deletions: Bystro returns all possible CADD Phred scores at such sites, whereas GEMINI returns a missing value. Bystro returns all values to give users added flexibility: its search engine can accurately search within arrays (lists) of data. Furthermore, as GEMINI on Galaxy only provided the Ensembl transcript set, for all query comparisons with GEMINI, Bystro was configured to use Ensembl 90, which was the latest version available at time of revision. It is important to note that the latest version of GEMINI on Galaxy (0.8.1) dates to February 2016, and its databases are several years older: CADD (v1.0, 2014), Ensembl (v75, February 2014), ExAc (v0.3, October 2014), whereas Bystro uses up-to-date resources. As a result of searching

more up to date Ensembl (v90), population allele frequency (gnomAD 2.0.1, the successor to ExAc 1.0), and CADD (v1.3) data, Bystro's queries returned more data. Since Galaxy does not report run times, QuickTime software was used to record each run, and the query time was calculated as the difference between the time the search submission entered the Galaxy queue, to the time that it was marked completed. Galaxy/GEMINI queries were each run more than 6 times. Because run times varied by more than 17x, the fastest consecutive 6 runs were averaged to minimize the influence of Galaxy server load.

All comparisons with the Bystro search engine are limited, because no other existing method provides natural-language parsing, and either rely on built-in scripts or require the user to learn a specific language (SQL).

Filtering accuracy comparison

The latest version of Bystro (beta 10, September 2017) was used. For the 1000 Genomes query accuracy checks, the same underlying Ensembl-based Bystro annotation and search index was used as in the Bystro/GEMINI filtering comparison. Direct comparison to GEMINI were not made, in reflection of the age of the latest GEMINI Galaxy version (v0.8.1, with database sources dating to 2014). All Bystro queries from that comparison were saved, downloaded, and compared with Bystro "Filters", which are exact-match alternatives to Bystro's natural-language queries, as well as custom Perl filtering scripts that also require exact matches. A second query accuracy step was conducted, on the Yen et al 2017 VCF file. This file was annotated using the standard RefSeq Bystro database. The same queries used in the

Bystro/GEMINI comparison were re-created on this smaller annotation, saved, downloaded, and compared with Bystro “Filters” and Excel filters. Excel filters were created in Excel 2016 (Mac), and required exact matches. All Excel-filtered and all Bystro query results were manually inspected for concordance (Additional file 2.7). All scripts generated and used in the comparison may be found at https://github.com/akotlar/bystro-paper.

Offline annotation comparisons

To generate offline performance data, the latest versions of each program available at time of writing were used. Bystro beta10 (September 2017), VEP 86 (March 2017), and ANNOVAR (March 2017) were each run on separate, dedicated Amazon i3.2xlarge EC2 instances (Additional file 2.3). All programs’ databases were updated to the latest versions available as of March 2017 (VEP, ANNOVAR), or September 2017 (Bystro). All programs were configured to use the RefSeq transcript set.

Each instance contained 4 CPU cores (8 threads), 60GB RAM, and a 1920GB NVMe SSD. Each instance was identically configured. All programs were configured to as closely match Bystro’s output as possible, although Bystro output more total annotation fields (Additional file 2.3). Each data set tested was run 3 times. The annotation time for each run was recorded, and averaged to generate the mean variant per second (variant/s) performance. Submissions were recorded using the terminal recorder asciinema, and both memory and cpu usage were recorded using the **free** and **top** commands set to a 30 second timeout.

VEP was configured to use 8 threads and to run in “offline” mode to maximize performance, as recommended (McLaren et al., 2016). In each of three recorded trials, VEP was set to annotate from RefSeq and CADD, and to check the reference assembly (Additional file 2.3). Based on VEP’s observed performance, adding PhastCons annotations was not attempted. VEP’s performance was measured by reading the program’s log, which records variant/second performance every $5x10^3$ annotated sites. In consideration of time, VEP was stopped after at least $2x10^5$ variants were completed, and the $2x10^5$ variants performance was recorded.

ANNOVAR was configured to annotate RefSeq, CADD, PhastCons 100way, PhyloP 100way, Clinvar, avSNP, and ExAc version 0.3 (Additional file 2.3). ANNOVAR’s avSNP database was used in place of dbSNP, as recommended. We configured ANNOVAR to report allele frequencies from ExAc, because it does not do so from either avSNP or dbSNP databases. When annotating Phase 1, Phase 3, or Phase 3 chromosome 1, ANNOVAR crashed by exceeding the available 60GB of memory. It was therefore tested with the subsets of Phase 3 chromosome 1 that contained $1x10^6$ – $4x10^6$ variants.

Bystro was configured to annotate descriptions from RefSeq, dbSNP 147, CADD, PhastCons 100way, PhyloP 100way, Clinvar, and to check the reference for each submitted genomic position (Additional file 2.3).

Annotation accuracy comparison

The latest version of Bystro (beta 10, September 2017), ANNOVAR (July 2017), and VEP (version 90) at the time of revision submission were used. All programs’

databases were updated to the latest version available. RefSeq-based databases were downloaded using each program's database builder. All programs were compared on the Yen et al. 2017 VCF file for position, variant call, and variant effects, based on each programs' respective RefSeq database. The Yen et al VCF file *fileformat* header line was modified to "VCFv4.1" to allow programs to recognize it as a valid VCF file. This modified file is available: https://github.com/akotlar/bystro-paper. For the SnpEff comparison, annotations were adapted from Additional File 1 of Yen et al. 2017. ANNOVAR was additionally configured with gnomAD genomes, gnomAD exomes, and CADD 1.3, and compared to Bystro on the corresponding values.

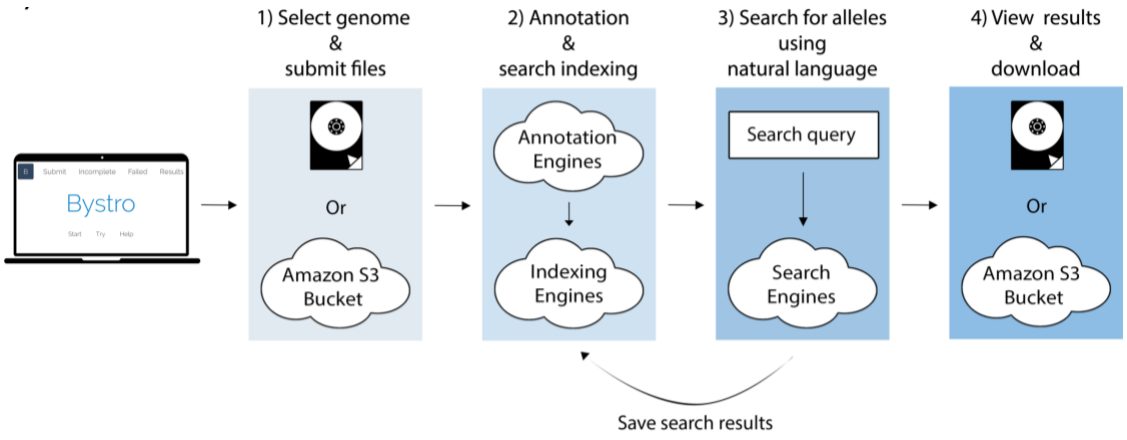


Figure 4.1 A | **Bystro use overview .** After logging in (https://bystro.io/), users upload one of more VCF or SNP-format files - containing alleles from a sequencing experiment - from a computer or a connected Amazon S3 bucket. Datasets of over 890GB, containing thousands of samples and tens of millions of variants are supported. The data is rapidly annotated in the cloud, using descriptions from public sources (e.g. RefSeq, dbSNP, Clinvar, and others). The annotated results can be filtered using Bystro's natural-language search engine, and any search results can be saved as new annotations. Annotated experiments and saved results can be viewed online, downloaded as tab-delimited text, or uploaded back to linked Amazon S3 buckets.

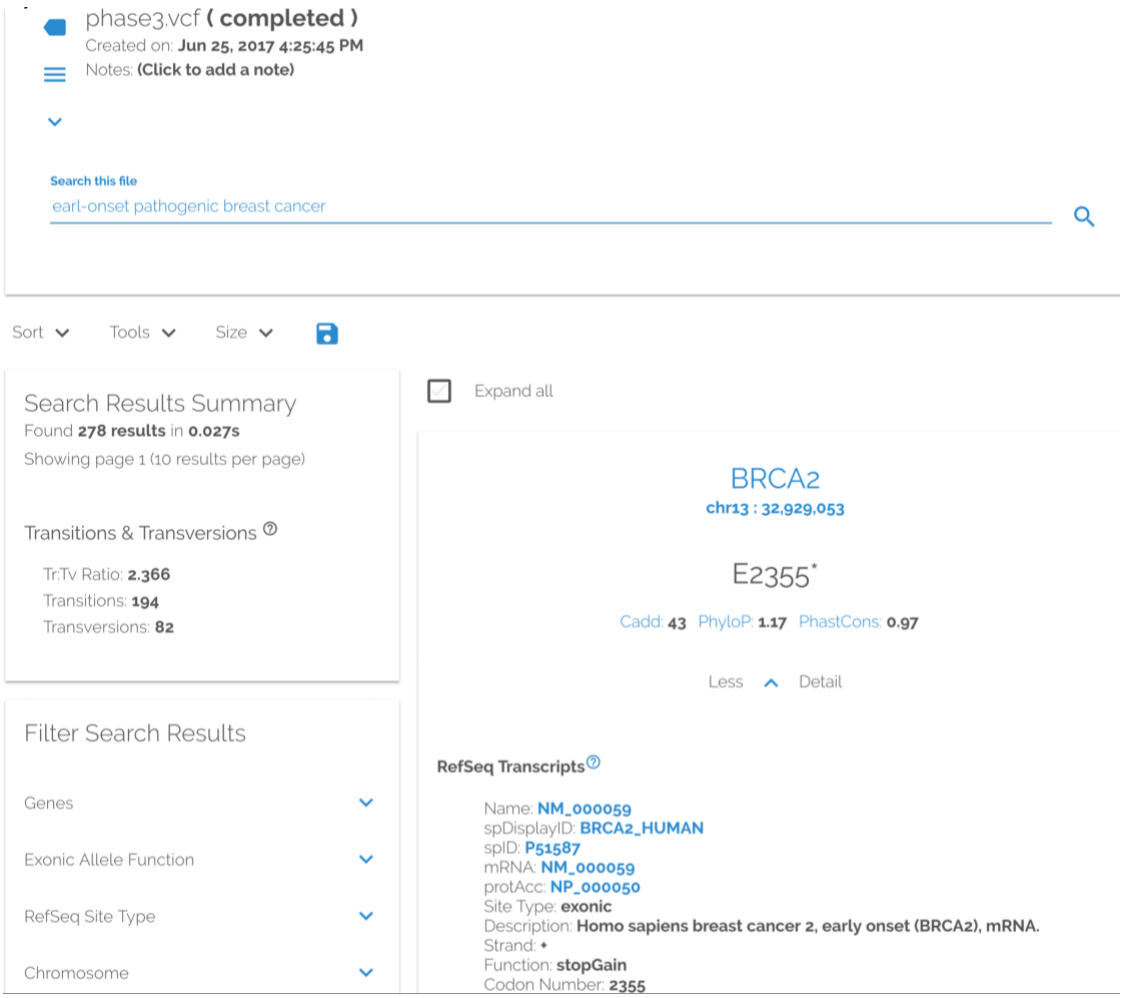


Figure 4.1 B | **Variant selection using Bystro.** An example of using Bystro's natural-language search engine to filter 1000 Genomes Phase 3 (https://bystro.io/public). To do so, users may type natural phrases, specific terms, numerical ranges, or apply filters on any annotated field. Queries are flexible, allowing misspelled terms such as "earl-onset" to accurately match. Complex tasks, such as identifying *de novo* variants can be achieved by using Boolean operators (AND, OR, NOT, +, -), exact-match filters, and user-defined terms. For instance, after labeling the "proband" and their "parents", the user could simply search *proband –parents*, or combine with additional parameters for more refined queries, i.e. *proband –parents missingness < .1 gnomad.exomes.af_nfe <*

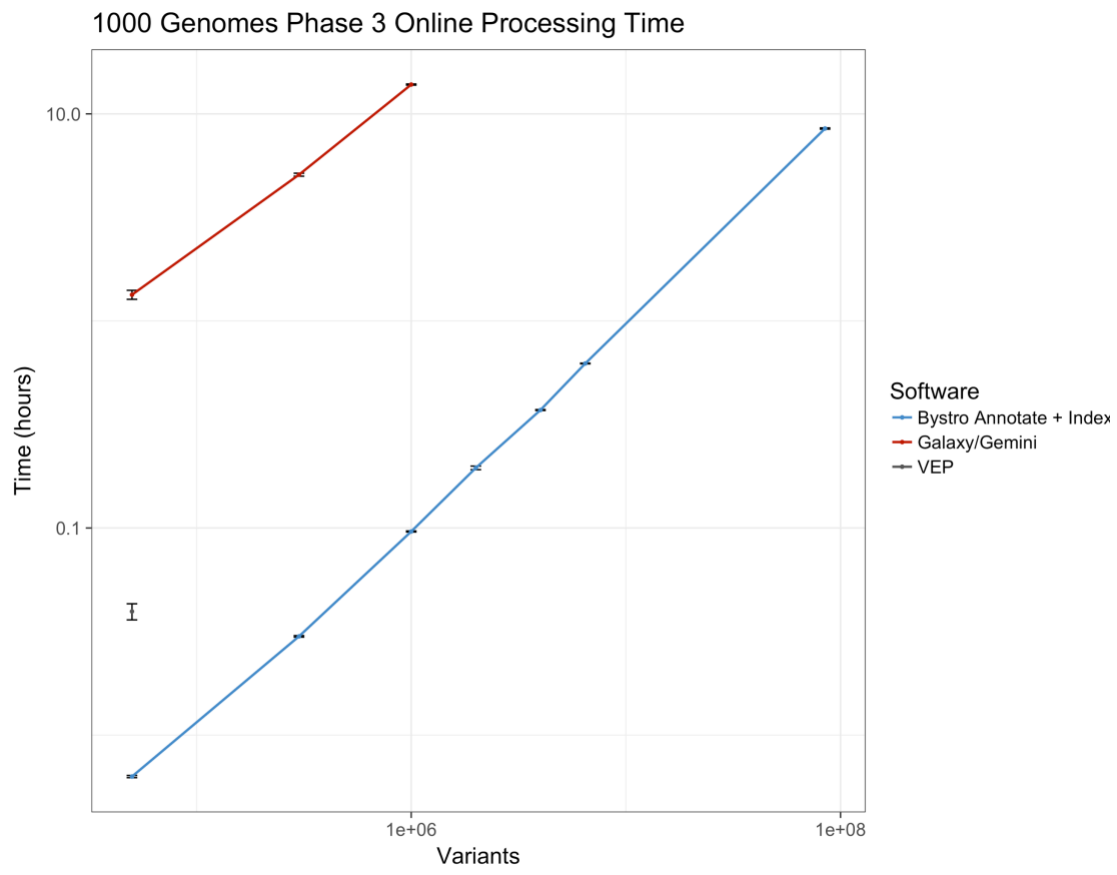


Figure 4.2 | **Online performance comparison of Bystro, VEP, wANNOVAR, and GEMINI**. Bystro, wANNOVAR, VEP, and GEMINI (running on Galaxy) we run under similar conditions. Total processing time was recorded for 1000 Genomes Phase 3 WGS VCF files, containing either the full data set (2,504 samples, 8.49×10^7 variant sites), or subsets (2,504 samples and 5×10^4, 3×10^5, 1×10^6, and 6×10^6 variants). Only Bystro successfully processed more than 1×10^6 variants online: wANNOVAR (not shown) could not complete the smallest 5×10^4 variant subset; VEP could not complete more than 5×10^4 variants; and GEMINI/Galaxy could not complete more than 1×10^6 variants. GEMINI and Bystro (but not VEP) outputted whole-genome CADD scores, while only Bystro also returned whole-genome PhyloP and PhastCons conservation scores. Bystro was faster than GEMINI/Galaxy by 144x-212x across all

Table 4.1 | **Bystro, VEP, ANNOVAR offline command-line performance**.

Software	Dataset	Samples	Variants	Variants/s	vs Bystro
Bystro	**1kG Phase 3 ch1**	**2504**	**1x10$_6$**	**8156 ± 195**	-
	1kG Phase 3 ch1	**2504**	**2x10$_6$**	**8484 ± 67.9**	-
	1kG Phase 3 ch1	**2504**	**4x10$_6$**	**8516 ± 57.2**	-
	1kG Phase 3 ch1	**2504**	**6.5x10$_6$**	**7779 ± 21.8**	-
	1kG Phase 1	**1092**	**3.9x10$_7$**	**5417 ± 76.8**	
	1kG Phase 3	**2504**	**8.5x10$_7$**	**7904 ± 15.9**	-
VEP	**1kG** Phase 1	1092	3.9x10$_7$	18.67 ± 0.58	290x
	1kG Phase 3	2504	8.5x10$_7$	10.00 ± 0.00	790x
ANNOVAR	**1kG** Phase 3 ch1	2504	1x10$_6$	74.67 ± 0.21	109x
	1kG Phase 3 ch1	2504	2x10$_6$	75.32 ± 0.06	113x
	1kG Phase 3 ch1	2504	4x10$_6$	75.15 ± 0.39	113x
	1kG Phase 3 ch1	2504	6.5x10$_6$	NA	NA
	1kG Phase 1	1092	3.9x10$_7$	NA	NA
	1kG Phase 3	2504	8.5x10$_7$	NA	NA

Bystro, VEP, and ANNOVAR were similarly configured with 8 threads on Amazon i3.2xlarge servers. “Dataset” refers to the VCF file used. “Variants/s” is the number of variants annotated per second, averaged across three trials. VEP performance was recorded after 2x10$_5$ sites in consideration of time. In runs of 1x10$_6$ or more annotated sites, VEP performance did not deviate from the 2x10$_5$ value. ANNOVAR could not complete the full Phase 1, Phase 3, or Phase 3 chromosome 1 datasets due to memory limitations. Thus, ANNOVAR was compared to Bystro on subsets of 1000 Genomes Phase 3 chromosome 1. Bystro run times included time taken to compress outputs. 1000 Genomes Phase 1 performance reflects IO limitations.

Table 4.2 | **Online comparison of Bystro and recent programs in filtering 8.49×10^7 variants from 1000 Genomes**

Group	Search query	Time (s)	Variants	Ts/Tv
1	exonic	0.03 ± 0.03	993,343	2.96
2 (a)	cadd > 20 maf < .001 pathogenic expert review missense	0.03 ± 0.01	65	1.71
2 (b)	cadd > 20 maf < .001 pathogenic **expert's** review **non-synonymous**	0.04 ± 0.02	65	1.71
2 (c)	cadd > 20 maf < .001 **pathogen** expert-**reviewed nonsynonymous**	0.04 ± 0.03	65	1.71
3 (a)	early onset breast cancer	0.05 ± 0.03	4,335	2.51
3 (b)	**early-onset** breast cancer	0.04 ± 0.02	4,335	2.51
3 (c)	**Early onset** breast **cancers**	0.03 ± 0.02	4,335	2.51
4 (a)	Pathogenic nonsense Ehlers-Danlos	0.04 ± 0.03	1	NA
4 (b)	**pathogenic** nonsense **E.D.S**	0.08 ± 0.09	1	NA
4 (c)	**pathogenic stopgain eds**	0.04 ± 0.02	1	NA

The full 1000 Genomes Phase 3 VCF file (853GB, 8.49×10^7 variants, 2,504 samples) was filtered in the publicly-available Bystro web application using the Bystro natural-language search engine. VEP, GEMINI, and wANNOVAR (not shown) were also tested, but were unable to annotate this data set or filter it. Bystro's search engine uses a natural language parser that allows for unstructured queries: queries in groups 2, 3, and 4 show phrasing variations that did not affect results returned, as would be expected for a search engine that could handle normal language variation. "Ts/Tv" is the transition to transversion ratio automatically calculated for each query by the search engine. The transition to transversion ratio of 2.96 for the "exonic" query is close to the ~2.8-3.0 ratio expected in coding regions, suggesting that the search engine accurately identified exonic (coding) variants.

References

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., … National Eye Institute, N. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Chang, X., & Wang, K. (2012). wANNOVAR: Annotating genetic variants for personal genomes via the web. *Journal of Medical Genetics*, *49*(7), 433–436. https://doi.org/10.1136/jmedgenet-2012-100918

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

DeFreitas, T., Saddiki, H., & Flaherty, P. (2016). GEMINI: A computationally-efficient search engine for large gene expression datasets. *BMC Bioinformatics*, *17*, 102. https://doi.org/10.1186/s12859-016-0934-8

Ganna, A., Genovese, G., Howrigan, D. P., Byrnes, A., Kurki, M., Zekavat, S. M., Whelan, C. W., Kals, M., Nivard, M. G., Bloemendal, A., Bloom, J. M., Goldstein, J. I., Poterba, T., Seed, C., Handsaker, R. E., Natarajan, P., Mägi, R., Gage, D., Robinson, E. B., … Neale, B. M. (2016). Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. *Nature Neuroscience*, *19*(12), 1563–1565. https://doi.org/10.1038/nn.4404

Goecks, J., Nekrutenko, A., Taylor, J., & Galaxy Team. (2010). Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, *11*(8), R86. https://doi.org/10.1186/gb-2010-11-8-r86

Johnston, H. R., Chopra, P., Wingo, T. S., Patel, V., International Consortium on Brain and Behavior in 22q11.2 Deletion Syndrome, Epstein, M. P., Mulle, J. G., Warren, S. T., Zwick, M. E., & Cutler, D. J. (2017). PEMapper and PECaller provide a simplified approach to whole-genome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(10), E1923–E1932. https://doi.org/10.1073/pnas.1618065114

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., & Maglott, D. R. (2016). ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, *44*(D1), D862–D868. https://doi.org/10.1093/nar/gkv1222

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., … Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association

mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, *20*(1), 110–121. https://doi.org/10.1101/gr.097857.109

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS Computational Biology*, *9*(10), e1003285. https://doi.org/10.1371/journal.pcbi.1003285

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. https://doi.org/10.1093/nar/29.1.308

Shetty, A. C., Athri, P., Mondal, K., Horner, V. L., Steinberg, K. M., Patel, V., Caspary, T., Cutler, D. J., & Zwick, M. E. (2010). SeqAnt: A web service to rapidly identify

and annotate DNA sequence variations. *BMC Bioinformatics*, *11*, 471. https://doi.org/10.1186/1471-2105-11-471

Taylor, R. C. (2010). An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinformatics*, *11 Suppl 12*, S1. https://doi.org/10.1186/1471-2105-11-S12-S1

Yen, J. L., Garcia, S., Montana, A., Harris, J., Chervitz, S., Morra, M., West, J., Chen, R., & Church, D. M. (2017). A variant by any name: Quantifying annotation discordance across tools and clinical databases. *Genome Medicine*, *9*(1), 7. https://doi.org/10.1186/s13073-016-0396-7

V. **Discussion**

V.I Conclusions

Modern methods in genomics have allowed us to explore the genetics of complex or rare incompletely penetrant disorders in different ways than the original concept of GWAS studies and to develop tools using the information gathered from these studies to prioritize candidates (Basu & Pan, 2011; Seunggeun Lee et al., 2012; Seunggeung Lee et al., 2014; Pan et al., 2015). GWAS studies over the past decade, specifically those primarily focused on common diseases, have revealed over 70,000 variant-trait associations, but there still remains a great deal of variation associated with many rare diseases that has yet to be explained (Buniello et al., 2019). Approaches that group variants into genes or pathways that take advantage of improved annotation of the genome give us a better understanding of overall burden that can be attributed to certain disorders (Seunggeun Lee et al., 2012). In addition, various study designs, combined with the new genetic tools, can be used to maximize the potential to identify risk variants. These include studying a complex trait in a genetically-sensitized population or establishing case and control definitions of a complex trait using the tails of the phenotypic distribution. In the work described here, we studied two separate disorders in sensitized populations - POI in PM women (FXPOI) and AVSD in DS (DS-AVSD).

In the study of modifying genes associated with FXPOI, candidate genes were prioritized through a genomic-analysis pipeline including SKAT-O, followed by a whole organism functional study using a *Drosophila* model. This strategy identified two

candidate genes that appear to have a synergistic effect with the fragile X premutation (PM). These have not previously been associated with idiopathic POI or FXPOI - *KRR1* and *SUMO1*. A PRS study of this cohort also identified approximately 7% of the variability between women with the PM with and without POI could be attributed to variants associated with age at natural menopause. These analyses taken together show that these methods are useful to find complex trait associations in small sample sizes, in this case 65 cases and 51 controls and can be used to generate and screen candidates in FXPOI. Confirmation and further functional studies are now needed.

For the study of DS-CHD, a total cohort of 702 individuals with DS and with or without AVSD was analyzed using a similar strategy to examine genetic risk factors associated with DS-AVSD. This study took advantage of available WES and WGS data and, given the inability to combine these data, developed a strategy to follow the WES findings with the candidate-gene approach using the WGS data. Using this approach, two top gene candidates were found - *NOTCH4* and *CEP290*, and their corresponding pathways were also found to be associated. The Notch pathway and the ciliome have previously been identified in playing a role in DS-associated AVSD, and these findings corroborate those results. The first PRS for AVSD was also done in this study. There was suggestive evidence that a PRS derived from associations with CHD in general explained some of the risk for DS-AVSD, with ORs ranging from 1.2 to 1.3 and corresponding Nagelkerke's r_2 values of approximately 1% (adjusted p-values > 0.15).

V.II Limitations

For these studies, two main genomic analyses were implemented - SKAT-O and PRS. For these types of analyses, variants are grouped into sets to either evaluate the burden of certain genes or pathways and overall polygenic risk, respectively. Being able to determine to which genes or pathways these variants belong, as well as knowing the corresponding tissues in which the genes are expressed, require these data to be annotated by previous work. One difficulty for determining whether genes are expressed in certain tissues is that gene expression information in humans is widely known for adults (*GTEx Portal*) but expression data for developmental stages is mostly limited to animal models. All genes found in genomic studies that are filtered through the annotations that are currently available may not include some of these genes, which is a limitation. As these data are determined by future studies and annotation methods make it easier to filter and sort through data quickly and effectively, more information will be revealed about complex traits. As more genomic studies start to include more complete WGS data instead of data derived by older sequencing methods, more rare variants will also be revealed in their association with complex disorders.

In the whole-organism functional assay for the SKAT-O results of the FXPOI study, top candidate genes were limited to genes that had an ortholog in *Drosophila*, so many genes that may be involved in ovarian dysfunction were not tested in a functional assay. Those genes could be further tested in cell culture or in mouse studies. Another limitation from the fecundity study is that given its purpose as a reporter of ovarian dysfunction, mechanism of any gene knockdown tested is difficult to determine from this assay alone and requires further work to understand the interaction between the 90 CGG repeat and knockdown of the candidate genes. Individual mutations that were

found in the women with the PM were also not tested during this study, and could be an avenue for further study in the PM mice.

V.III Implications and future directions

Results from both studies form the foundation for confirmation studies and for specific hypothesis testing (e.g., further investigation of the Notch pathway). At this point, our results are too preliminary to suggest any clinical applications (e.g., screening women who carry a premutation for variants in *SUMO1).* Future directions for these studies include looking at confirmation cohorts of larger sample sets. In addition, it would be interesting to ask whether the identified variants are associated with the full spectrum of the disorder, not just the extremes. For example, a study of age at diagnosis of FXPOI/age at menopause among PM carriers will now be important to understand how much of the variation is explained. Similarly for the candidate genes associated with DS-AVSD, it is now important to ask whether these same genes are associated with the other forms of CHD associated with DS. Next, the study of these genes in cohorts that include individuals with idiopathic POI and nonsyndromic AVSD are warranted. Such studies may begin to identify subgroups of individuals with these particular disrupted pathways.

Follow-up studies to understand the functional role of candidate genes is essential. Our use of the *Drosophila* PM model was simply to provide a secondary screen of highly ranked genes. Follow-up studies in mammalian systems can be used to understand mechanism. Thus, going forward with the results of the FXPOI study,

determining the mechanism of the top candidate genes from the analysis using one of the mouse models is the immediate next step. To follow up on the DS-CHD results, examining the top results of the analysis with a functional assay (e.g., zebrafish) and then conducting further studies of mechanism in the mouse model for DS are required.

The approaches used in this dissertation that combined the statistical approaches of SKAT-O and PRS with a study design involving a genetically-sensitized population and using extreme phenotypes, has helped to maximize the limited sample size to identify candidate genetic risk factors. The use of a model organism as a whole-organism functional assay to screen these candidates helped to prioritize research on the disease mechanisms. Such approaches can be applied to other complex disorders to further understand their genetic architecture and determine the potential to translate such findings to the clinical are.

References

Basu, S., & Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology*, *35*(7), 606–619. https://doi.org/10.1002/gepi.20609

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(Database issue), D1005–D1012.

https://doi.org/10.1093/nar/gky1120

GTEx Portal. (n.d.). Retrieved June 13, 2020, from https://gtexportal.org/home/documentationPage#staticTextPublicationPolicy

Lee, Seunggeun, Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., NHLBI GO Exome Sequencing Project—ESP Lung Project Team, Christiani, D. C., Wurfel, M. M., & Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*, *91*(2), 224–237. https://doi.org/10.1016/j.ajhg.2012.06.007

Lee, Seunggeung, Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-Variant Association Analysis: Study Designs and Statistical Tests. *American Journal of Human Genetics*, *95*(1), 5–23. https://doi.org/10.1016/j.ajhg.2014.06.009

Pan, W., Chen, Y.-M., & Wei, P. (2015). Testing for polygenic effects in genome-wide association studies. *Genetic Epidemiology*, *39*(4), 306–316. https://doi.org/10.1002/gepi.21899