*Database*, 2020, 1–9
doi:10.1093/database/baaa083
Database update

Database update

# Color Data v2: a user-friendly, open-access database with hereditary cancer and hereditary cardiovascular conditions datasets

**Mark J. Berger[#], Hannah E. Williams[#], Ryan Barrett, Anjali D. Zimmer, Wendy McKennon, Huy Hong, Jeremy Ginsberg, Alicia Y. Zhou and Cynthia L. Neben[*]**

Color Genomics, 831 Mitten Road, Suite 100, Burlingame, CA, 94010, USA

[*]Corresponding author: Tel: +1 (650) 651-7116, Email: clneben@color.com

[#]These authors contributed equally to this work.

Citation details: Berger,M.J., Williams,H.E., Barrett,R. *et al.* Color Data v2: a user-friendly, open-access database with hereditary cancer and hereditary cardiovascular conditions datasets. *Database* (2020) Vol. XXXX: article ID baaa083; doi:10.1093/database/baaa083

Received 15 April 2020; Revised 27 August 2020; Accepted for publication 2 September 2020

## Abstract

Publicly available genetic databases promote data sharing and fuel scientific discoveries for the prevention, treatment and management of disease. In 2018, we built Color Data, a user-friendly, open access database containing genotypic and self-reported phenotypic information from 50 000 individuals who were sequenced for 30 genes associated with hereditary cancer. In a continued effort to promote access to these types of data, we launched Color Data v2, an updated version of the Color Data database. This new release includes additional clinical genetic testing results from more than 18 000 individuals who were sequenced for 30 genes associated with hereditary cardiovascular conditions as well as polygenic risk scores for breast cancer, coronary artery disease and atrial fibrillation. In addition, we used self-reported phenotypic information to implement the following four clinical risk models: Gail Model for 5-year risk of breast cancer, Claus Model for lifetime risk of breast cancer, simple office-based Framingham Coronary Heart Disease Risk Score for 10-year risk of coronary heart disease and CHARGE-AF simple score for 5-year risk of atrial fibrillation. These new features and capabilities are highlighted through two sample queries in the database. We hope that the broad dissemination of these data will help researchers continue to explore genotype–phenotype correlations and identify novel variants for functional analysis, enabling scientific discoveries in the field of population genomics.

**Database URL:** https://data.color.com/

## Introduction

The use of next-generation sequencing (NGS) technologies in research and clinical laboratories has led to a rapid increase of genetic data. However, there is a lack of publicly available genetic data, especially paired genotypic–phenotypic data. In 2018, we launched Color Data, an open-access, cloud-based database containing genotypic and self-reported phenotypic information from 50 000 individuals who were sequenced for 30 genes associated with hereditary cancer (1). Color Data has already made an impact on the scientific community, being utilized in peer-reviewed publications (2, 3) and presented as a resource to educators (4). Its user-friendly interface enables researchers to easily execute their own queries with self-serve filters and displays the results as text, tables and graphs. Results can also be downloaded in different file formats for further analyses and shared via email or social media.

Another important consideration when designing and implementing the database was scalability for volume and integration of different data points. For the second release of Color Data ('Color Data v2'), we added new features to the existing hereditary cancer dataset as well as a new dataset of genotypic and self-reported phenotypic information related to hereditary cardiovascular conditions from more than 18 000 individuals. Importantly, the hereditary cardiovascular conditions dataset retains the same user-friendly interface, making it easy for researchers and scientists to explore a new disease area.

Here we describe updates made to the database, including changes to the cohort and the addition of new query filters and results such as family health history. We also added clinical risk models to Color Data v2: the Gail Model (5) and the Claus Model (6) for breast cancer, simple office-based Framingham Coronary Heart Disease Risk Score (7) for coronary heart disease and CHARGE-AF simple score (8) for atrial fibrillation. These risk models are commonly used by healthcare providers in the clinic and are important tools to estimate risk. Recent work has demonstrated that polygenic scores can also accurately predict and stratify risk for common, complex diseases and can identify individuals who have magnitude of risk for disease similar to those with pathogenic or likely pathogenic variants (9, 10). As such, Color Data v2 also includes polygenic scores for breast cancer, coronary artery disease and atrial fibrillation. We highlight the addition of the hereditary cardiovascular conditions dataset, clinical risk models and polygenic scores through two sample queries in the database. To our knowledge, Color Data is the first database to include pre-calculated scores from clinical risk models and for polygenic risk, which can help researchers investigate the relationship between different types of risk factors, both genetic and non-genetic, for disease.

## Materials and methods

Design and implementation of the database were previously described in the flagship publication by Barrett, Neben *et al.* (1). All individuals included in Color Data v2 received a multi-gene NGS panel test from Color Genomics, Inc. ('Color', Burlingame, CA) for 30 genes associated with hereditary cancer. In addition, a subset of individuals also received multi-gene NGS panel testing for 30 genes associated with hereditary cardiovascular conditions. All individuals consented to have their genetic and self-reported phenotypic information appear in Color's research database.

Laboratory procedures, bioinformatics analysis and variant interpretation for the multi-gene panel tests were performed at Color (Burlingame, CA) under Clinical Laboratory Improvements Amendments (#05D2081492) and College of American Pathologists (#8975161) compliance as previously described (11). Bioinformatics analysis included the previously described 30 genes associated with hereditary cancer and was updated to include an additional 30 genes associated with hereditary cardiovascular conditions: *ACTA2, ACTC1, APOB, COL3A1, DSC2, DSG2, DSP, FBN1, GLA, KCNH2, KCNQ1, LDLR, LMNA, MYBPC3, MYH7, MYH11, MYL2, MYL3, PCSK9, PKP2, PRKAG2, RYR2, SCN5A, SMAD3, TGFBR1, TGFBR2, TMEM43, TNNI3, TNNT2* and *TPM1*. Variant classification and classification categories used in the database reflect those in use at the time of database release. Variant classification categories for Color Data v2 are pathogenic, likely pathogenic, variant of uncertain significance (VUS), likely benign, and benign. There are several alleles that are classified as pathogenic or likely pathogenic by multiple submitters in ClinVar but are known in the field to be commonly occurring and of low penetrance, specifically *APC* c.3920T > A (p.I1307K), *CHEK2* c.470T > C (p.I157T), *MITF* c.952G > A (p.E318K) and all heterozygous pathogenic or likely pathogenic variants in *MUTYH* for hereditary cancer. These alleles would have been reported as positive results and are included in the pathogenic frequency calculation in the hereditary cancer dashboard. Analysis, variant calling and reporting focused on the complete coding sequence and adjacent intronic sequence of the primary transcript(s), unless otherwise indicated. In *APOB*, exon 1 was not analyzed, and VUS were not reported. In *MYH7*, variants of uncertain significance were not reported for exon 27. In several genes, certain exons were not analyzed: exons 4 and 14 of *KCNH2,* exon 1 of *KCNQ1,* exon 11 of *MYBPC3,* exon 5 of *PRKAG2* and exon 1 of *TGFBR1*.

Laboratory procedures and imputation for low coverage whole genome sequencing were performed at Color as previously described (12, 13). Data from low coverage

whole genome sequencing were used to calculate previously published polygenic scores for breast cancer (10), coronary artery disease (9) and atrial fibrillation (9). Each polygenic score was normalized using principal components analysis to account for the effects of population stratification. While polygenic scores have the highest performance in people of European ancestry, recent studies have demonstrated that they have stratification ability across global populations as well (14, 15). To note, if users would like to view polygenic risk score results for a given query, they must select 'Calculated' in the polygenic risk score filter because only a subset of the individuals in the database have a calculated polygenic risk score. Individuals who do not have polygenic risk scores calculated are captured under the filter value 'Unknown'. Other self-reported phenotypic and genotypic information from 'Calculated' and 'Unknown' individuals is included in other query results by default, unless otherwise selected.

Genotypic and self-reported phenotypic information were used in the following clinical risk models: Gail Model for 5-year risk of breast cancer (5), Claus Model for lifetime risk of breast cancer (6), simple office-based Framingham Coronary Heart Disease Risk Score for 10-year risk of coronary heart disease (7) and CHARGE-AF simple score for 5-year risk of atrial fibrillation (8). To note, only a subset of individuals have a risk score calculated. Individuals who do not have a risk score calculated are labeled as 'Unknown' if not enough information was provided to calculate a risk score or 'Ineligible' if they did not meet the model criteria. The eligibility criteria for risk models are as follows:

- Gail Model: female, age 35 to 85 years, no personal history of breast cancer, no likely pathogenic or pathogenic variants in a gene associated with hereditary breast cancer (*BRCA1*, *BRCA2*, *TP53*, *PTEN*, *STK11*, *CDH1*, *PALB2*, *CHEK2*, *ATM*, *NBN*, *BARD1* and *BRIP1*)
- Claus Model: female, age 30 to 79 years, no personal history of breast cancer, no likely pathogenic or pathogenic variants in a gene associated with hereditary breast cancer (*BRCA1*, *BRCA2*, *TP53*, *PTEN*, *STK11*, *CDH1*, *PALB2*, *CHEK2*, *ATM*, *NBN*, *BARD1* and *BRIP1*)
- Simple office-based Framingham Coronary Heart Disease Risk Score: age 30 to 74 years
- CHARGE-AF simple score: age 46 to 90 years

Risk score filter values for hereditary cancer are defined in Table 1 and for hereditary cardiovascular conditions in Table 2.

## Results

### Web interface

The Color Data home page (https://data.color.com/) links to two new query/result pages (hereafter referred to as 'dashboards'): one for hereditary cancer and one for hereditary cardiovascular conditions. The links to three sample queries on the home page have been updated to demonstrate to users potential use cases of these dashboards as well as new query filters and results.

On the hereditary cancer dashboard (https://data.color.com/v2/cancer.html), users can apply the new query filters for family health history, risk models and a polygenic risk score. These new filter categories and filter values are listed in Table 1. To note, the 'AND' logic for filter categories and 'OR' logic for filter values within categories still apply. New query results for hereditary cancer include 'Gail Risk Score—5-Year Risk of Breast Cancer', 'Claus Risk Score—Lifetime Risk of Breast Cancer' and 'Breast Cancer (BC) Polygenic Risk Score'. For 'Breast Cancer (BC) Polygenic Risk Score', results will only be displayed if a user selects the filter value 'Calculated' because only a subset of individuals have a polygenic risk score calculated.

On the hereditary cardiovascular conditions dashboard (https://data.color.com/v2/cardio.html), users can apply the same types of query filters as those available on the hereditary cancer dashboard, with the following substitutions for clinical risk models and polygenic risk scores: 'CHARGE—AF Risk Score', 'Framingham Risk Score', 'AF Polygenic Risk Score' and 'CAD Polygenic Risk Score'. For 'AF Polygenic Risk Score' and 'CAD Polygenic Risk Score', results

**Table 1.** New filter categories and filter values on the hereditary cancer dashboard.

| Filter categories | Filter values |
|---|---|
| Gail Risk Score | Elevated risk (age 35–49 years and risk ≥1.67%; age 50–59 years and risk ≥2.0%; age 60–69 years and risk ≥3.0%; age 70–74 years and risk ≥4.0%), Ineligible*, Non-elevated risk, Unknown† |
| Claus Risk Score | Elevated (≥20%), Ineligible*, Non-elevated (<20%), Unknown† |
| BC Polygenic Risk Score‡ | Calculated, Unknown† |

*Ineligible includes females who did not meet the model criteria and males.
†Unknown includes females who did not provide enough information to calculate risk scores.
‡Males included. Only displays normalized risk scores between − 3 and 3.

**Table 2.** Filter categories and filter values on hereditary cardiovascular conditions dashboard.

| Filter categories | Filter values |
|---|---|
| Sex | Female, Male |
| Age | 18–25, 26–30, 31–35, 36–40, 41–45, 46–50, 51–55, 56–60, 61–65, 66–70, 71–75, 76–80, 81–85, 86–89, ≥90 |
| Ethnicity | African, Ashkenazi Jewish, Asian, not specified; Caucasian, Chinese, Filipino, Hispanic, Indian, Japanese, Multiple ethnicities, Native American, Pacific Islander, Unknown* |
| Personal Health History | Aneurysm, Angioplasty, Bundle branch or heart block, Bypass, Cardiac arrest, Cardiomegaly, Heart attack, Heart failure, No cardiac events, Stent, Stroke |
| Family Health History | Aneurysm, Angioplasty, Bundle branch or heart block, Bypass, Cardiac arrest, Cardiomegaly, Heart attack, Heart failure, Stent, Stroke |
| Classification | Benign, Likely Benign, Likely Pathogenic, Pathogenic, VUS |
| Gene | *ACTA2, ACTC1, APOB, COL3A1, DSC2, DSG2, DSP, FBN1, GLA, KCNH2, KCNQ1, LDLR, LMNA, MYBPC3, MYH7, MYH11, MYL2, MYL3, PCSK9, PKP2, PRKAG2, RYR2, SCN5A, SMAD3, TGFBR1, TGFBR2, TMEM43, TNNI3, TNNT2, TPM1* |
| Variant | (Search by HGVS Nomenclature)† |
| Zygosity | Heterozygous, Homozygous |
| CHARGE-AF Risk Score | Borderline (5%–7.5%), High (>10%), Ineligible‡, Intermediate (7.5%–10%), Low (<5%), Unknown§ |
| Framingham Risk Score | Borderline (5%–7.5%), High (>10%), Ineligible‡, Intermediate (7.5%–10%), Low (<5%), Unknown§ |
| Polygenic Risk Score**φ** | Calculated, Unknown |

*Unknown includes information not reported.
†Filter values for 'Variant' can only be selected by text typing with auto complete using Human Genome Variation Society (HGVS) nomenclature.
‡CHARGE-AF ineligible includes individuals <46 or ≥90 years. Framingham ineligibility includes individuals <30 or >74 years.
§Unknown includes individuals who did provide enough information to calculate risk scores.
**φ** Only displays normalized risk scores between − 3 and 3.

will only be displayed if a user selects the filter value 'Calculated' because only a subset of individuals have polygenic risk scores calculated. These filter categories and filter values are listed in Table 2. The same 'AND' logic and 'OR' logic apply, as described above.

## Population characteristics

The population characteristics of the hereditary cancer dataset in Color Data v2 are very similar to those in Color Data v1. However, due to changes in inclusion criteria, there are some notable differences. These include an increase in the proportion of men (27.6% versus 20.4%) and non-Caucasian individuals (30.6% versus 27.9%). The frequency of pathogenic and likely pathogenic variants in the total population increased to 11.0% in Color Data v2, compared with 10.8% in Color Data v1 (note: pathogenic frequency includes low common penetrance alleles such as *APC* c.3920T > A [p.I1307K], *CHEK2* c.470-T > C [p.I157T], *MITF* c.952G > A [p.E318K] and all heterozygous pathogenic or likely pathogenic variants in *MUTYH*). There are a total of 14 269 unique variants in Color Data v2. The newly added query results show that 9.3% of individuals have an elevated, 5-year risk of breast cancer as estimated using the Gail Model, and only 0.8% have an elevated lifetime risk of breast cancer as estimated using the Claus Model.

The 18 783 individuals in the new hereditary cardiovascular conditions dataset are a subset of the individuals in the hereditary cancer dataset. The majority of individuals are female (68.5%) and reported Caucasian ethnicity (70.4%). The average age at the time of genetic testing was 46.2 years. A total of 14 213 (75.6%) individuals reported no personal history of cardiovascular disease and/or events. Approximately 190 (1.0%) individuals reported a personal history of having a stroke, and 170 (0.9%) individuals reported having a heart attack. A total of 2 464 482 variants were identified in 30 genes associated with hereditary cardiovascular conditions, with the largest percentages in *RYR2, MYH11* and *DSP*. There are 9727 unique variants, over half of which are benign or likely benign (54.4%). The frequency of pathogenic and likely pathogenic variants in the total population is 1.4%. Finally, 0.3% of individuals are categorized as being at high-risk for atrial fibrillation using the CHARGE-AF simple score, and 7.5% are estimated to have a high 10-year risk for coronary heart disease using the simple office-based Framingham Coronary Heart Disease Risk Score.

## Sample query 1: frequency of pathogenic and likely pathogenic variants in genes associated with hereditary cardiovascular conditions

Cardiovascular disease is a leading cause of death in the USA, accounting for one-third of deaths worldwide

(16). Many individuals with hereditary cardiovascular conditions progress asymptomatically, and as a result, go undiagnosed until they present with a sudden cardiac event. Users can investigate the frequency of pathogenic and likely pathogenic variants in genes associated with hereditary cardiovascular conditions in the database by filtering 'Classification: Pathogenic or Likely pathogenic' (https://data.color.com/v2/cardio.html#classification=Likely%20pathogenic&classification=Pathogenic). A total of 223 individuals have a pathogenic or likely pathogenic variant, the majority of which are female (67.7%) (Figure 1A) and Caucasian (76.2%) (Figure 1B). The average age at testing was 45.0 years (Figure 1A), and the majority of individuals reported no personal history of cardiovascular disease and/or events (66.8%) (Figure 1C). Nearly one-fourth of variants were identified in *LDLR* (19.6%), followed by *MYBPC3* (18.1%), *KCNQ1* (9.7%) and *PKP2* (9.7%), among others (Figure 1D). Of the 223 pathogenic or likely pathogenic variants identified, the most common result was a heterozygous *APOB* c.10580G > A (p. Arg3527Gln) (n = 16) (Figure 1E), which is associated with familial hypercholesterolemia and found in approximately 0.06% of individuals of European, non-Finnish ancestry (17, 18). Familial hypercholesterolemia is characterized by elevated levels of low-density lipoprotein (LDL) and an increased risk of premature coronary artery disease, with 50% of men and 30% of women developing coronary artery disease by the age of 55, if left untreated (19). To investigate the subpopulation of individuals with this variant, users can filter by 'Gene: APOB' and 'Variant: c.10580G>A' (https://data.color.com/v2/cardio.html#gene=APOB&variant=c.1058-0G%3EA). In this subpopulation of 16 individuals, the majority were female (68.8%) and of Caucasian ethnicity (87.5%) (Figure 1F). On average, individuals in this subpopulation were 45.2 years old at the time of testing, and one individual reported a personal history of bundle branch block or heart block (Figure 1G). Taken together, researchers could use these data to better characterize the prevalence of hereditary cardiovascular disorders in a younger, unaffected population to identify asymptomatic individuals who are at risk for future cardiovascular disease and/or events.

### Sample query 2: monogenic and polygenic breast cancer risk in women with a personal history of breast cancer

Breast cancer is a common, complex disease that is associated with rare pathogenic and likely pathogenic variants ('monogenic risk') and the cumulative effect of many common changes across the genome ('polygenic risk') (9, 10). Recent work has suggested that monogenic risk and polygenic risk interact to modify an individual's overall risk for disease (13, 20–22). To investigate monogenic and polygenic risk in women with a personal history of breast cancer, users can filter by 'Sex: Female', 'Personal health history: Breast' and 'BC Polygenic Risk Score: Calculated' (https://data.color.com/v2/cancer.html#sex=Female&personal_health_history=Breast&bc_polygenic_risk_score=Calculated). A total of 1443 females in the dataset reported a personal history of breast cancer and had a polygenic risk score for breast cancer risk calculated (Figure 2A). The majority of individuals are Caucasian (73.9%) (Figure 2B), with an average age of 59.9 years at the time of genetic testing (Figure 2A). The average age of diagnosis for breast cancer was 53.2 ± 11.2 years (standard deviation), and 39.2% of females were <50 years old at the time of diagnosis (Figure 2C). The pathogenic frequency in this population was 13.3% (Figure 2A). A total of 82 677 variants were identified, with the majority of variants in *BRCA2* (14.6%), *BRCA1* (12.0%) and *APC* (11.3%) (Figure 2D). Compared with the normal distribution of risk scores among all individuals with a Polygenic Score, the distribution in women with a personal history of breast cancer (individuals filtered by query) is left-skewed (Figure 2E). Taken together, users could use these data to investigate the risk conferred by monogenic and polygenic risk factors in women with a history of breast cancer.

## Discussion

In Color Data v2, we added new query filters and results such as family health history as well as clinical risk models and a polygenic score for breast cancer to the existing hereditary cancer dataset. Overall, the total number of individuals in this dataset increased to 54 000; however, the individuals in Color Data v1 are not a strict subset of the individuals in Color Data v2. This is due to a difference in inclusion criteria between the two versions. In Color Data v1, individuals were included in the database if they had received clinical genetic testing for all or any subset of the 30 hereditary cancer genes. In Color Data v2, individuals were only included if they received genetic testing for all 30 genes. This change in cohort composition likely contributed to the observed change in frequency of pathogenic and likely pathogenic variants in the total population and the increase in the number of total and unique variants.

Database users can also now explore a new disease area with self-reported phenotypic information and genetic data for 30 genes associated with hereditary cardiovascular conditions from 18 738 individuals. The frequency of pathogenic and likely pathogenic variants

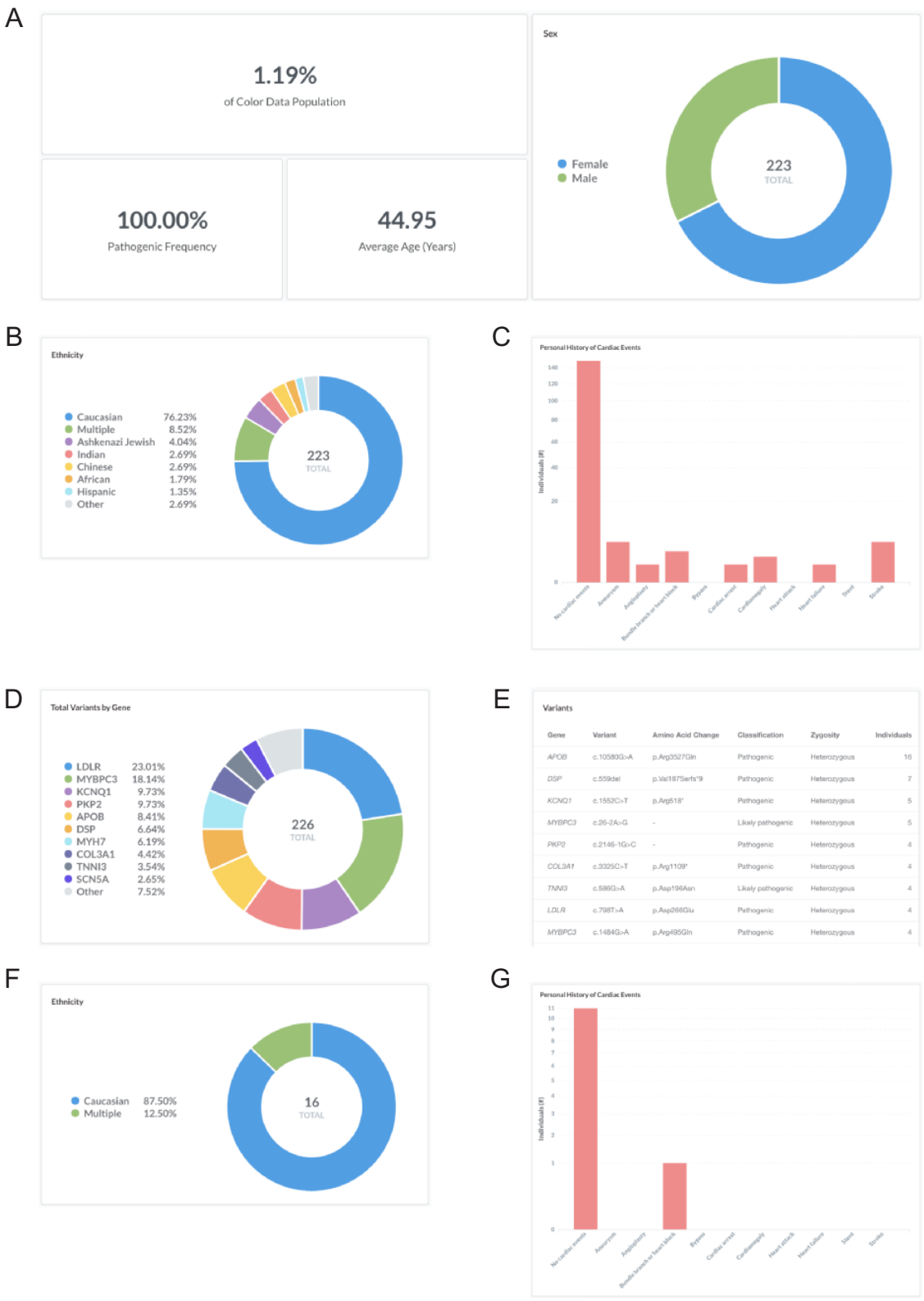


**Figure 1.** Screenshots of query results for frequency of pathogenic and likely pathogenic variants in genes associated with hereditary cardiovascular conditions. (A–E) Filter by 'Classification: Pathogenic or Likely pathogenic'. Query URL: https://data.color.com/v2/cardio.html#classification=Likely%20pathogenic&classification=Pathogenic (F, G) Remove 'Classification: Pathogenic or Likely pathogenic' and filter by 'Gene: APOB' and 'Variant: c.10580G>A'. Query URL: https://data.color.com/v2/cardio.html#gene=APOB&variant=c.10580G%3EA.

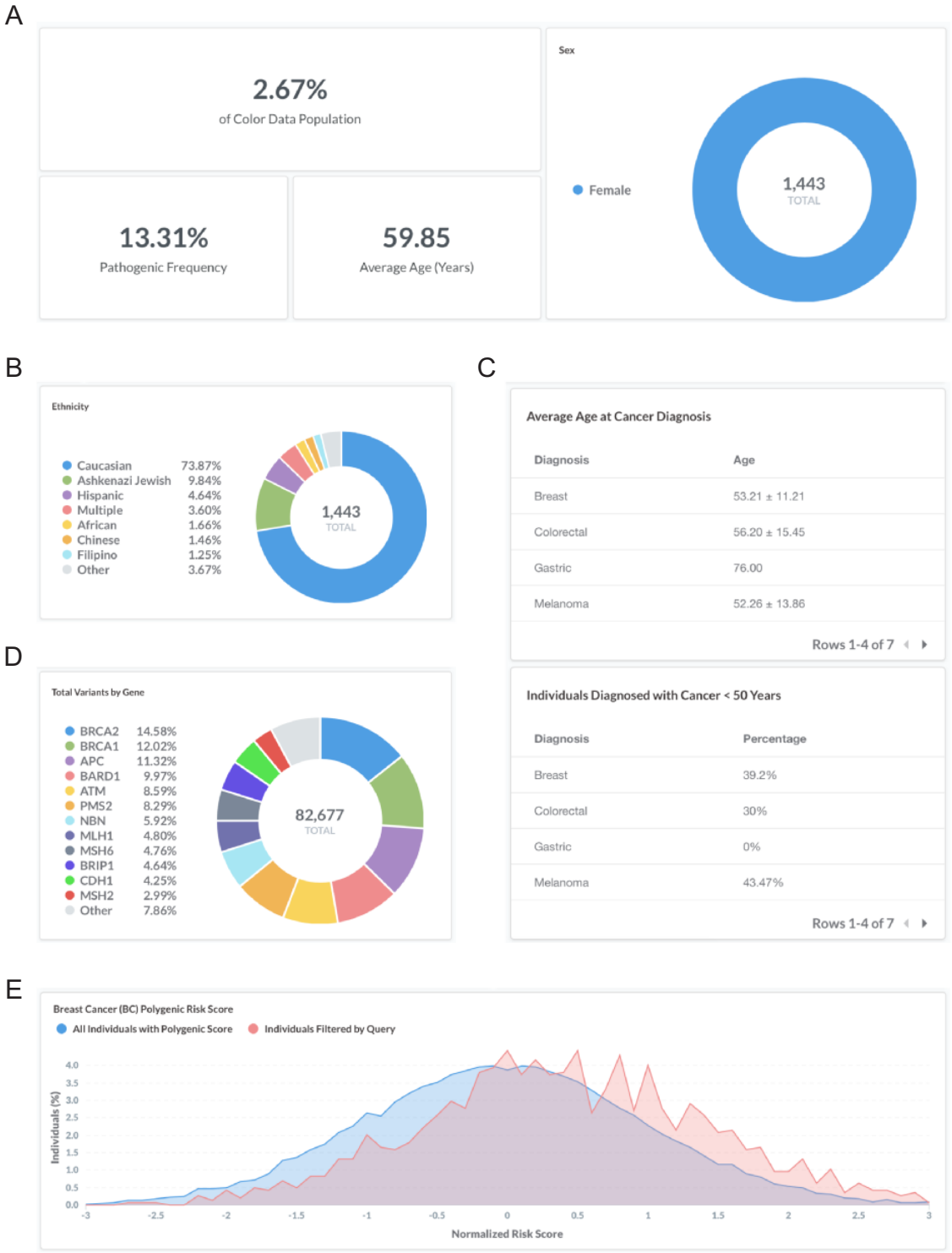


**Figure 2.** Screenshots of query results for monogenic and polygenic breast cancer risk in women with a personal history of breast cancer. (A–E) Filter by 'Sex: Female', 'Personal health history: Breast' and 'BC Polygenic Risk Score: Calculated'. Query URL: https://data.color.com/v2/cancer.html#sex=Female&personal_health_history=Breast&bc_polygenic_risk_score=Calculated.

in our dataset was higher than previously reported estimates (23, 24). This could be due to the generally younger age of individuals in the cohort and/or reduced penetrance in asymptomatic carriers. Genetic testing for hereditary cardiovascular conditions at population scale has only recently begun, and sharing results through genetic databases such as Color Data will help rapidly advance our understanding of cardiovascular disease risk. Coronary

artery disease may be of particular interest given the influence lifestyle modifications have been shown to have on lowering risk for disease. In a prospective study of more than 55 000 individuals, it was found that a healthy lifestyle was associated with significantly reduced risk of cardiovascular events across all genetic risk groups (25).

Similar to Color Data v1, Color Data v2 may be limited by selection bias for Caucasians and women as well as by self-reported phenotypic information. Not all individuals in the database provided enough information to calculate risk for the newly included clinical risk models or had polygenic risk scores calculated, resulting in incomplete datasets. As the field continues to evaluate the personal and clinical utility of polygenic risk scores, it will be important to consider their predictive power in light of other risk factors. In addition, the clinical risk models and polygenic scores shown may change over time as more evidence emerges and novel models are developed.

## Acknowledgements

We would like to thank Carman Lai and Sydney Okumura for insightful user testing and Michael K. Doney, Julian R. Homburger, Lauren Ryan and Stephanie E. Wallace for helpful discussions.

## Funding

This work was supported by Color Genomics.

*Conflicts of interest* All authors are currently employed and have equity interest in Color Genomics. R.B. was previously employed at Google. J.G. was previously employed at Twitter.

## Data sharing statement

The data in this report are publicly available at Color Data (https://data.color.com/). All reported variants have been submitted to ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/submitters/505849/).

## References

1. Barrett, R., Neben, C.L., Zimmer, A.D. *et al.* (2019) A scalable, aggregated genotypic-phenotypic database for human disease variation. *Database*, **2019**.10.1093/database/baz013.
2. Ndugga-Kabuye, M.K. and Issaka, R.B. (2019) Inequities in multi-gene hereditary cancer testing: lower diagnostic yield and higher VUS rate in individuals who identify as Hispanic, African or Asian and Pacific Islander as compared to European. *Fam. Cancer*, **18**, 465–469.
3. Kwon, D.H.-M., Borno, H.T., Cheng, H.H. *et al.* (2019) Ethnic disparities among men with prostate cancer undergoing germline testing. *Urol. Oncol.*, **38**, 80.e1–80.e7.
4. (2018) Science Extension | Garvan institute of medical research https://www.garvan.org.au/research/kinghorn-centre-for-clinical-genomics/learn-about-genomics/for-teachers/extension-science (accessed Jan 13, 2020).
5. Gail, M.H., Brinton, L.A., Byar, D.P. *et al.* (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.*, **81**, 1879–1886.
6. Claus, E.B., Risch, N. and Thompson, W.D. (1994) Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. *Cancer*, **73**, 643–651.
7. D'Agostino, R.B.Sr, Vasan, R.S., Pencina, M.J. *et al.* (2008) General cardiovascular risk profile for use in primary care: the Framingham heart study. *Circulation*, **117**, 743–753.
8. Alonso, A., Krijthe, B.P., Aspelund, T. *et al.* (2013) Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. *J. Am. Heart Assoc.*, **2**, e000102.
9. Khera, A.V., Chaffin, M., Aragam, K.G. *et al.* (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.*, **50**, 1219–1224.
10. Mavaddat, N., Michailidou, K., Dennis, J. *et al.* (2019) Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Hum. Genet.*, **104**, 21–34.
11. Neben, C.L., Zimmer, A.D., Stedden, W. *et al.* (2019) Multi-Gene panel testing of 23,179 individuals for hereditary cancer risk identifies pathogenic variant carriers missed by current genetic testing guidelines. *J. Mol. Diagn.*, **21**, 646–657.
12. Homburger, J.R., Neben, C.L., Mishne, G. *et al.* (2019) Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. *Genome Med.*, **11**, 74.
13. Fahed, A.C., Wang, M., Homburger, J.R. *et al.* (2020) Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat. Commun.*, **11**, 3635.
14. Khera, A.V., Chaffin, M., Zekavat, S.M. *et al.* (2019) Whole-genome sequencing to characterize monogenic and polygenic contributions in patients hospitalized with early-onset myocardial infarction. *Circulation*, **139**, 1593–1602.
15. Martin, A.R., Kanai, M., Kamatani, Y. *et al.* (2019) Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.*, **51**, 584–591.
16. Murphy, S.L., Xu, J., Kochanek, K.D. *et al.* (2018) Mortality in the United States, 2017. *NCHS Data Brief*, **328**, 1–8.
17. gnomAD. https://gnomad.broadinstitute.org/variant/2-2122-9160-C-T?dataset=gnomad_r2_1 (accessed Aug 6, 2020).
18. Fernández-Higuero, J.A., Etxebarria, A., Benito-Vicente, A. *et al.* (2015) Structural analysis of APOB variants, p.(Arg3527Gln), p.(Arg1164Thr) and p.(Gln4494del), causing Familial hypercholesterolaemia provides novel insights into variant pathogenicity. *Sci. Rep.*, **5**, 18184.
19. Slack, J. (1969) Risks of ischaemic heart-disease in familial hyperlipoproteinaemic states. *Lancet*, **2**, 1380–1382.
20. Fahed, A.C., Wang, M., Homburger, J.R. *et al.* (2019) Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. *Genetic Genomic Med.*, **11**.
21. Kuchenbaecker, K.B., McGuffog, L., Barrowdale, D. *et al.* (2017) Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. *J. Natl. Cancer Inst.*

22. Oetjens, M.T., Kelly, M.A., Sturm, A.C. *et al.* (2019) Quantifying the polygenic contribution to variable expressivity in eleven rare genetic disorders. *Nat. Commun.*, **10**, 4897.
23. Blanch, B., Sweeting, J., Semsarian, C. *et al.* (2017) Routinely collected health data to study inherited heart disease: a systematic review (2000-2016). *Open Heart*, **4**, e000686.
24. Semsarian, C., Ingles, J., Maron, M.S. *et al.* (2015) New perspectives on the prevalence of hypertrophic cardiomyopathy. *J. Am. Coll. Cardiol.*, **65**, 1249–1254.
25. Khera, A.V., Emdin, C.A., Drake, I. *et al.* (2016) Genetic risk, adherence to a healthy lifestyle, and coronary disease. *N. Engl. J. Med.*, **375**, 2349–2358.