



**Coriell Personalized Medicine Collaborative (CPMC®)
Technical Paper
V2.0 September 22, 2009**

Coriell Institute for Medical Research
403 Haddon Avenue
Camden, NJ 08103

Section 1. Introduction

Overall Goals

The Coriell Personalized Medicine Collaborative (CPMC[®]) is a multi-party effort involving study participants, medical professionals, scientists, ethicists and information technology experts. The overall goal of the study is to evaluate the utility of using the knowledge of genetics in medicine when implemented in an ethical, legal and responsible format. Using secure, web-based interfaces, participants complete extensive medical history, family history and lifestyle questionnaires and report their current medication use. Once completed, they are given access to personalized results for potentially actionable health conditions. The results are in the form of a risk report, where genetic variant risk and in some instances non-genetic risks are displayed. Participants are given the opportunity to share this information with their healthcare providers, family members and friends via the web portal. In the future, the CPMC[®] will report results on genetic variants that influence a person's response to commonly used medications in addition to risk reports for common complex diseases.

Following the release of results, participants are asked at regular intervals to determine whether receipt of this information influenced their personal health decisions; if they shared their results with their healthcare provider, their medical care; and ultimately their overall health. Annually, participants are asked to update their medical history, family history and lifestyle questionnaires. These data will be analyzed to determine if providing results for genetic variants that have been associated with complex diseases directly to research participants results in changes in participant health and/or health behavior. In addition, CPMC[®] scientists will analyze the genetic and medical data to validate previously identified genetic variants and identify new genetic variants which influence the risk of common complex diseases. CPMC[®] will also examine how multiple genetic variants and non-genetic risk factors can be used to estimate risk of disease.

Participants are given the option to release their de-identified data to the biomedical research community; with separate assent provided for not-for-profit and for-profit institutions. CPMC[®] will create a database of de-identified genetic and medical history information from participants who indicate that their data may be released in this way. The dataset will be deposited into a national database where it will be archived and distributed to researchers and access will be overseen by a data access committee. A data access committee is made up of scientists and staff at the National Institutes of Health. The committee reviews applications for access to and use of the de-identified data. Access is limited to those research uses of the data that comply with the limits set in the consent document. The research that may be performed using this data includes identification of genes and gene variants that elevate the risk of common complex diseases and/or alter response to particular medications.

Use of an External Advisory Board

All health conditions and genetic variants must be approved by an independent advisory board called the Informed Cohort Oversight Board (ICOB). This board is charged with assessing the validity of the associations between the proposed genetic variant and disease condition, and confirming that all proposed conditions are potentially actionable (i.e., a condition for which the disease risk can be

mitigated by individual action (behavior, lifestyle) or by medical action (screening, preventative treatment, early intervention)). Assessment of risk mitigation focuses on the disease itself, as in most instances, it is unknown if the risk associated with a genetic variant can be mitigated. The board votes on condition and genetic variant submissions as described below and the majority vote rules. The board includes scientists skilled in genetic studies, medical professionals familiar with the use of genetics in medical care, an ethicist as well as a pastor who can share insight into the moral and ethical implications of the study as well as provide valuable feedback on their understanding of genetic results as members of a lay audience. ICOB members are not employed by Coriell Institute for Medical Research. The ICOB meets at least twice a year to review conditions and genetic variants selected by the CPMC[®]. Following the meeting, the ICOB members review and approve meeting summaries generated by CPMC[®] staff. Deliberations of the ICOB are not made public; however, the voting results of the board will be posted on the CPMC[®] web portal.

Section 2. Selection of conditions and variants for ICOB nomination

Health Condition Selection

Coriell reviews health conditions for which genetic variants have been identified as risk alleles. Additionally, Coriell reviews the condition to determine if it is potentially actionable. An actionable condition is defined as a condition for which the risk can be mitigated by individual action (behavior, lifestyle) or by medical action (screening, preventative treatment, early intervention).

An example of a health condition that is unlikely to be actionable is amyotrophic lateral sclerosis (ALS). If prevention, screening and treatment for ALS changes significantly such that it may meet the “actionable” metric, the CPMC[®] may submit ALS to the ICOB for consideration.

Coriell reviews medical society policies and recommendations to assess if a health condition is potentially actionable as defined above. It compiles a Condition Summary document that is reviewed for accuracy by one or more CPMC[®] clinician advisor(s). The Condition Summary contains the following information:

- brief description of the condition
- incidence and prevalence data
- known causes of the condition
- prevention strategies
- screening procedures
- policy statements and guidelines, when available

The Condition Summary is submitted with a Genetic Variant Summary to the Informed Cohort Oversight Board (ICOB).

Genetic Variant Selection

Coriell reviews peer-reviewed scientific literature and websites including HuGENet (<http://www.cdc.gov/genomics/hugenet/default.htm>) and the NHGRI GWAS catalogue (<http://www.genome.gov/26525384>) to identify genetic variants that have been found to be associated with a health condition that Coriell wishes to submit to the ICOB for review. The genetic variant must be found to be associated with the disease in more than one cohort of the same race, either replicated within a single peer-reviewed publication or published in separate peer-reviewed

publications. The primary means of testing for genetic variants is via the Affymetrix 6.0 GeneChip. Affymetrix NetAffx (<http://www.affymetrix.com/analysis/index.affx>) is used to identify genetic variants contained on the chip via the dbSNP RefSNP (rs) ID, gene name and/or probe set ID. Only genetic variants contained on the Affymetrix 6.0 GeneChip will be submitted to the ICOB for consideration at this time. Custom SNP genotyping will be implemented with the plan to add additional genotyping platforms during the course of the study.

The Genetic Variant Summary contains the following information:

- gene name, chromosomal location, gene variant ID, probe set ID, alleles
- cohort size, cohort race/ethnicity, gene frequency, odds ratios, p-values, PMID of primary publications
- population genotype and allele frequencies in the 4 primary International HapMap Project populations (www.hapmap.org)
- Other association studies of this variant with other conditions
- Other association studies of variants in this gene region with other conditions
- Other variants in this gene region associated with this condition
- Variants in other genes associated with this condition
- Variants in other genes associated with this condition submitted to the ICOB

The Genetic Variant Summary is submitted with a Condition Summary to the Informed Cohort Oversight Board (ICOB).

ICOB Review of Submissions

The ICOB meets at least twice per year. The ICOB receives the submissions from the CPMC[®] at least two weeks prior to the ICOB meeting. The chair of the ICOB assigns primary and secondary reviewers to each condition and associated genetic variant submissions. The primary and secondary reviewers present a synopsis to the board. After discussion, the ICOB votes on whether they deem the condition potentially actionable. They can approve, approve with minor revision, defer, or not approve the condition. The vote on the condition is independent of the vote of the genetic variant, but the genetic variant is not voted on unless the condition receives a vote of "approve" or "approve with minor revision". Those genetic variants that are approved are processed by CPMC[®] scientists to generate risk reports, as described below. Those submissions not approved by the ICOB can be resubmitted in the future if Coriell finds that data published since the initial submission strengthens the evidence for actionability or genetic association.

Section 3. Risk Estimation

Genetic Variant Risk Determination

The following sections (Study Selection, Relative Risk Estimation, Estimates in Non-Caucasians, and Multigenic Models) describe the procedure Coriell follows to determine risk estimates between a specific disease condition and single genetic variant. Coriell presents risk estimates for both genetic variants and non-genetic risk factors as relative risk, derived or reported from a valid and representative publication. Relative risks are given instead of absolute risks (lifetime risk or 10-year risk as a %, for example) due to the lack of prospective studies reporting absolute risk from genetic, lifestyle and other factors, and the lack of studies providing estimates of absolute risk from models that account for competing risks and changes in risk-factor status over time.

Study Selection for Genetic Risk

Study selection consists of a number of steps. Coriell developed its study selection strategy by incorporating current published recommendations (Attia et al., 2009; Ioannidis et al., 2008). It considers peer-reviewed scientific literature in which ICOB-approved genetic variants and diseases are studied. Initially, Coriell considers the references used to support the ICOB approval including publications from general PubMed and HuGENet searches using the genetic variant and disease terms, and the papers they reference. Coriell's first step is to group these studies based upon their design and rank the designs based upon their expected ability to provide representative and valid estimates of association, using the hierarchy of study designs from Table 1.

Table 1. Hierarchy of Study Designs

	Study Design
Tier 1	Meta-Analysis of Prospective Studies
Tier 2	Individual Prospective Study
Tier 3	Meta-Analysis of Case-Control Studies
Tier 4	Individual Population-Based Case-Control Study
Tier 5	Individual Case-Control Study (not population based)

Once Coriell isolates studies from the highest available tier (smallest tier number) of the design hierarchy, it examines study quality and reporting ability. Study quality items that Coriell considers include the disease definition (was it reported, was it carried out consistently, and was it objective and clinically accepted), genotyping methods (was it reported, was it carried out consistently) and, when relevant, population stratification. Meta-analyses that only report results that pool across more than one ethnic group (and do not report results within each ethnic group) will only be used if the study report includes justification for the pooling. Coriell also assesses whether or not the publication reports estimates of risk. If none of the studies from the highest tier of study design meet these quality and reporting criteria, then Coriell considers studies from the next study design tier. If more than one study from the highest tier of study design meets these quality and reporting criteria, Coriell chooses one paper based upon the largest number of included studies (meta-analysis) or greatest number of total cases and total participants. However, if there is one prospective study in a U.S. population that provides estimates of both genetic and non-genetic risk estimates, then this study is given preference. The preference for U.S. population-based studies is due to possible differences in the impact of lifestyle and environmental factors on disease risk by country, and the assumption that estimates from U.S. populations would be most relevant to CPMC[®] participants.

Relative Risk Estimation

Once Coriell selects the study for risk reporting, it determines the risk estimates to present. The goal is to present risk estimates as relative risk. Relative risk (or odds ratios which estimate relative risk in some situations) are reported more consistently in the literature, provide a quantitative measure of

the association between the risk factor and disease and are generally comparable across populations with different underlying risk of disease. Relative risk can be estimated directly in prospective studies but not in case-control studies. Odds ratios can be estimated from case-control studies, but odds ratios are not always good estimates of relative risk (Davies et al., 1998; Jewell, 2003). When the underlying disease prevalence is relatively high (over about 10%) or the odds ratio is not close to the null value (larger or smaller than 1), the odds ratio may markedly overstate the relative risk. Under these circumstances, it may be misleading to interpret odds ratios as relative risk (see below for CPMC[®]'s approach to calculating relative risk from odds ratios).

Thus, when the selected study is a prospective study, estimates of relative risk are usually determined directly. If the prospective study reports estimates of relative risk (including hazard ratios from survival analyses), then those values of relative risk are reported. If the prospective study reports estimates of odds ratios and the prevalence of disease in the unexposed group in the population are available, then Coriell calculates relative risk, based upon the relationship between the relative risk, odds ratio and prevalence of disease in the unexposed group ($RR = OR / [(1 - p_0) + p_0 \times OR]$ where p_0 = prevalence of disease in the unexposed group). In rare situations when a prospective study reports estimates of absolute risk, Coriell reports absolute risk along with relative risk.

When the selected study is a case-control study, Coriell must estimate relative risk based upon the reported odds ratio. It first finds an estimate of the underlying disease prevalence in the population that produced the cases and controls. Given the reported odds ratio and estimate of underlying disease prevalence, Coriell estimates the degree to which the odds ratio overstates the relative risk (Davies et al., 1998). If the odds ratio overstates the relative risk by an estimated 10% or less, then Coriell reports the odds ratio as an estimate of relative risk. If the odds ratio overstates the relative risk by more than 10%, then a conservative estimate for relative risk, calculated based upon the estimated overall disease prevalence, is given.

When a participant already has a diagnosis of the disease under study, the genetic variant relative risk is provided for information only since the risk estimate is for developing the disease in individuals without the disease.

For gender-specific diseases, relative risk estimates are not provided to participants of the unaffected gender since they are not personally at risk of disease. However, genetic variant genotype results are provided for information.

Estimates for non-Caucasians

Coriell initially determines risk estimates in Caucasian populations. For some genetic variants for which data are available, Coriell presents separate risk estimates for other races and/or ethnicities. It selects the paper for risk reporting by evaluating the study quality of available publications in non-Caucasian populations. Coriell considers the same study quality items as it uses when comparing potential studies with the same design in Caucasians (i.e., disease definition, genotyping methods, population stratification, and ability to report risk estimates). In addition, it assesses study size and generally requires that the study include at least 200 cases (and more when the risk allele frequency is less than 0.2 or the odds ratio from the Caucasian population is less than 1.3). In studies that include more than one race or ethnicity, Coriell confirms that analyses account for population stratification. It also assesses the methods used to account for multiple testing in studies that include more than one genetic variant.

If no studies in non-Caucasian populations are available, or if the studies are of poor quality, or the association between the genetic variant and disease is not replicated, then Coriell presents the risk estimates from the Caucasian population to non-Caucasian participants “for information only”. Similarly, if the participant does not report his or her race and ethnicity then Coriell presents the risk estimates from the Caucasian population. These study limitations are highlighted in the limitations section of the Risk Results Report.

Multigenic Models

At this time, Coriell does not provide genetic risk estimates based upon more than one genetic variant. In the future, it will consider reporting estimates based upon published multigenic models using genetic variants approved by the ICOB.

Non-Genetic Risk Determination

Coriell presents risk estimates for non-genetic risk factors as relative risk, based upon a valid and representative publication, and initially applies these methods to studies of Caucasian populations. In the future, Coriell also will present risk estimates for non-genetic risk factors in non-Caucasians, when the data are available in the peer-reviewed literature.

Study Selection for Family History and Other Risks

Coriell presents risk estimates due to non-genetic factors that (1) are clinically accepted disease risks, and (2) can be assessed based upon information collected by the Demographic and Medical History, Family History, Lifestyle Questionnaire (MFLQ). Non-genetic risk factors include family history, behavioral and environmental factors and co-morbidities. It reviews the epidemiologic literature to determine the risk factors of disease that are consistently reported and clinically accepted, and then determine which factors can be assessed based upon responses to the MFLQ. Coriell selects a valid and representative study, following a similar strategy to that used to find the publication used for genetic risk estimates, and reports relative risk using the same methods used when reporting genetic risk. Since most known disease risk factors have been studied prospectively, Coriell generally limits its pool of potential studies to prospective studies (or meta-analyses of prospective studies) conducted in U.S. populations. It requires the study to use a disease definition that is consistent with that used in the genetic association study, along with commonly accepted definitions of risk factors. Preference is given to studies conducted in the U.S. as it is assumed these estimates of risk due to environmental and behavioral factors will be most relevant to the CPMC[®] cohort. When possible, Coriell selects a study that reports more than one risk factor or uses the same population from the genetic association study.

When Coriell cannot find a study that reports a quantitative risk estimate for a known disease risk factor collected in the MFLQ, it may provide a qualitative, text result. If the necessary information for family history risk or other risk factors are not reported by the participant, then the risk report does not provide a specific risk estimate but provides the relative risk range for the risk factor based on the participant’s demographic information. In addition, the participant is informed that specific risk results are not available because necessary information for the risk factor was marked as “unknown” or not provided on the MFLQ. At this time, the majority of reported risk estimates for non-genetic risks are derived from studies in Caucasian populations.

Section 4. Other Estimates and Limitations

Prevalence Estimates

When possible, Coriell reports disease prevalence estimates by gender, age and race/ethnicity in the “How Common” tab of the risk report. These estimates are generally representative of the U.S. population and come from reports by groups such as the CDC, NCHS and SEER, or come from peer-reviewed literature. When prevalence estimates are not available for all demographic factors (gender, age, and race-ethnicity), prevalence estimates are provided using as many demographic factors as are available. Coriell provides a summary plot of the prevalence estimates to demonstrate how prevalence of disease varies by demographic factors. It also provides an estimate of disease prevalence with the risk report so as to help put the relative risks due to genetic and non-genetic factors within the context of the underlying population risk.

Heritability

Coriell reports heritability of disease estimates based upon values given in peer-reviewed publications. Generally, these studies are based upon twin or family data and have moderate sample size.

Genotype Frequency

Coriell reports genotype frequency by race/ethnicity, when such estimates are available, based upon the participant’s self-reported race and ancestry. If genotype frequencies are reported for the study population in which the genetic risk estimates were determined, then Coriell reports these values, unless they are statistically different ($p < 0.05$) from HapMap genotype frequencies in the same ethnic group. Otherwise, if another peer-reviewed publication reports genotype frequencies, for the specific ethnic group, in a population that is larger than HapMap and these values do not differ statistically from the HapMap frequencies, then Coriell reports these values. Otherwise, Coriell provides genotype frequencies from HapMap samples. For ethnic groups that are not included in HapMap, another ethnic group is selected as the default, and frequencies from the default group are reported (See Appendix Table 1 for details.). The population used is referenced in the “More Information” tab of the risk report. [Genotype frequencies are reported as percentages rounded to the nearest whole number \(i.e., 23% rather than 23.2%\).](#)

Formatting and Presentation Rules for Risk Estimates

Coriell reports relative risk using one decimal place and does not report estimates of variability, such as 95% confidence bounds, with risk estimates, due to the general target audience and amount of information it presents. It describes relative risk (n.n) using language such as “individuals with this genetic variant have n.n times the risk of developing the disease as those with no copies of the risk variant.” For relative risks less than 2, it also describes as “individuals with this genetic variant are at p% increased/decreased risk of developing the disease as compared to those with no copies of the risk variant”, where $p\% = (RR - 1) \times 100\%$ increased risk for $RR > 1$ and $p\% = (1 - RR) \times 100\%$ decreased risk for $RR < 1$. To avoid confusion for study participants, relative risks greater than 2 are not presented in terms of percent increase as this would be read as a greater than 100% increase in risk. Coriell presents absolute risk estimates as percentages.

Limitations

At this time, validated predictive models based upon well-studied, prospective populations with known genetic and non-genetic risk factors have not been developed, and so predicted risks of

disease based upon an individual's specific risk factors cannot be estimated. Instead, Coriell gives relative risks which are estimates based upon populations with a specific risk factor. Relative risks are not usually estimated using the same statistical model and population, and do not always come from models that adjust for other important factors. When only case-control studies are available, relative risks must be estimated based upon the reported odds ratio.

Summary

Coriell has designed a system to select genetic variants that have been identified and validated by the research community as associated with human disease. This system involves reviewing and monitoring the peer-reviewed literature to identify candidate genetic variants. It involves the use of national health organizations guidelines as well as clinical advisors to identify potentially actionable conditions. Most uniquely, it involves the use of an external committee of scientists, medical professionals, ethicists and religious leaders to review and vote on candidates for inclusion in the study. Those that are approved for inclusion are presented to study participants in the form of genotype and genetic risk along with non-genetic risk using a web-based reporting system.

Strengths of the CPMC[®] genetic variant selection and reporting system include the criteria that the variant be validated in the peer-reviewed literature and that candidates are vetted by an external oversight board. The strength of the risk report is that it includes non-genetic risk factors that are customized for the participant based on self-reported demographics, medical history, family history and lifestyle information. The participant can use this additional non-genetic risk information to put their genetic risk in perspective and have the opportunity to take action in areas where they can mitigate risk, such as in lifestyle choices like smoking. When data is available, the risk report includes risk due to family history. Emphasis of the role of family history in disease risk and health management is a positive outcome of participation in this study and may help individuals review their family history with their family members and communicate this information to their healthcare provider.

Risk estimates reported by the CPMC[®] come directly from peer-reviewed publications and do not require the application of broad assumptions. They provide participants with a view of the current state of the scientific literature and the need for continued research. The lack of large prospective studies examining genetic and non-genetic risks, and the lack of validated, predictive models developed in such populations limit what the CPMC[®] can report. The CPMC[®] cohort will provide valuable prospective data to advance this area of research.

When possible, risk information is based on datasets that match the race/ethnicity of the individual. The CPMC[®] is limited by the data available in the public domain; however, it is likely that the number of population datasets will grow in the future, and include more non-Caucasian populations. Study participants will learn that the applicability of genetic risk data to their ethnic group depends heavily on members of the ethnic group participating in genetic studies aimed at identifying genetic associations. This may heighten awareness about the value of minority participation in genetic research.

Complex diseases are caused by the interaction of multiple genes and environmental factors. The scientific community continues to identify new genes associated with disease. Thus, it is not possible currently, to completely measure an individual's genetic risk of a complex disease. Coriell is currently releasing individual genetic variant results and disease risk based on single genetic variants, while

stressing through its web content and genetic counseling that these results do not represent an individual's total genetic risk and do not address how genetic factors may interact with non-genetic factors. There is an effort underway in some common complex diseases to model risk based on multiple genetic factors. Coriell will integrate genetic risk based on multiple genetic factors when such risk models have been published and validated in the peer-reviewed literature. The CPMC[®] cohort can be very valuable in testing these algorithms, as the prospective study design will allow Coriell researchers to determine if the risk models are good predictors of health outcomes.

Coriell has designed the CPMC[®] study around the use of a secure web portal for information exchange between the study and the participant. The reporting of genetic results and risk using a visual display with access to extensive explanatory and educational material with a mechanism for gauging understanding and uptake through surveys will be extremely useful when designing the reporting of complex genetic results to medical professionals and integrating such information into electronic health records.

References

Attia, J., Ioannidis, J.P., Thakkinstian, A., McEvoy, M., Scott, R.J., Minelli, C., Thompson, J., Infante-Rivard, C., and Guyatt, G. (2009). How to use an article about genetic association: B: Are the results of the study valid? *Jama* *301*, 191-197.

Davies, H.T., Crombie, I.K., and Tavakoli, M. (1998). When can odds ratios mislead? *BMJ (Clinical research ed)* *316*, 989-991.

Ioannidis, J.P., Boffetta, P., Little, J., O'Brien, T.R., Uitterlinden, A.G., Vineis, P., Balding, D.J., Chokkalingam, A., Dolan, S.M., Flanders, W.D., *et al.* (2008). Assessment of cumulative evidence on genetic associations: interim guidelines. *International journal of epidemiology* *37*, 120-132.

Jewell, N.P. (2003). *Statistics for Epidemiology*.

Appendix Table 1
 CPMC[®] Reported Genotype Frequencies: Sources used by
 Ethnic Group

	White	Black	Hispanic	Native American	Pacific Islander	Asian	Asian_Ch	Asian_In	Asian_J	Asian_Ko	Mixed Race	Unknown
1st choice	Use the study used for genetic risk estimates (if this ethnic group included and genotype frequencies reported)*											
2nd choice	Use another paper reporting genotype frequencies in a large population of this ethnic group*											
3rd choice	HapMap CEU	1.HapMap ASW 2.HapMap YRI	<i>Default to Caucasian values</i>	<i>Default to Caucasian values</i>	<i>Default to Caucasian values</i>	<i>Report Asian-Ch values</i>	HapMap CHB	1.HapMap GIH 2.Default to Caucasian values	HapMap JPT	<i>Report Asian-Ch values</i>	<i>Default to Caucasian values</i>	<i>Default to Caucasian values</i>

*for studies in White, Black, or Asian groups, use estimates from publication as long as they do not statistically differ ($p < 0.05$) from HapMap estimates

HapMap Populations:

ASW: African ancestry in Southwest USA

CEU: Utah residents with Northern and Western European ancestry from the CEPH collection

CHB: Han Chinese in Beijing, China

GIH: Gujarati Indians in Houston, Texas

JPT: Japanese in Tokyo, Japan

YRI: Yoruba in Ibadan, Nigeria