Supplementary Information for Systematic Design and Comparison of Expanded Carrier Screening Panels

Kyle A. Beauchamp, PhD*1 Dale Muzzey, PhD*1 Kenny K. Wong, MS CGC*1
Gregory J. Hogan, PhD1 Kambiz Karimi, PhD1 Sophie I. Candille, PhD1
Nikita Mehta, MS CGC1,2 Rebecca Mar-Heyming, PhD1 K. Eerik Kaseniit, MEng1
H. Peter Kang, MD1 Eric A. Evans, PhD1 James D. Goldberg, MD1,4
Gabriel A. Lazarin, MS CGC1 Imran S. Haque, PhD1,3

* These authors contributed equally to this work
1 Counsyl. 180 Kimball Way, South San Francisco, CA 94080. (888) 268-6795
2 Current address: Mayo Clinic, Rochester, MN
3 Current address: Freenome, South San Francisco, CA
4 Corresponding author: research@counsyl.com

Contents

I Supplementary Materials and Methods 2
  1 CNV Deletion Calling 2
  2 Census Weighting 2
  3 Disease Risk Calculation 2
    3.1 Additive Approximation to Disease Risk 2
    3.2 Example: Autosomal Recessive Disease Risk for one Disease 3
    3.3 Calculating Disease Risk 3
    3.4 Designing Optimal Targeted Genotyping Panels 3

II Supplementary Tables and Figures 5
  5 Supplementary Table S1 6
  6 Supplementary Figure S1 7
  7 Supplementary Table S2 8
  8 Supplementary Figure S2 9
  9 Supplementary Figure S3 10
  10 Supplementary Table S3 11
  11 Supplementary Figure S4 12
  12 Supplementary Table S4 13
  13 Supplementary Table S5 14
  14 Supplementary Figure S5 15
Part I
Supplementary Materials and Methods

1 CNV Deletion Calling

Most expanded carrier screening tests provide copy number variation (CNV) calling for a limited set of genes. To explore the impact of full, panel-wide CNV deletion calling, we applied a simple CNV calling algorithm to identify putative CNV calls. In preparation for CNV analysis, the read counts for each region were normalized to account for a sample’s overall sequencing depth, capture and mapping efficiency in the region, and relative GC bias. After normalization, samples or specific regions with high residual noise were censored. For CNV calling, a sliding window approach identified regions with putative deletions or duplications based on a z-score statistic. Only deletions were considered pathogenic during subsequent analysis.

Panel-wide CNV analysis was skipped for several genes; some of these genes are special cases (CYP21A2, FMR1, HBA1, SMN1, GBA), while others (Table S3; e.g., CTNS) have specific CNV analyses already performed as part of Counsyl Family Prep Screen. For the genes with pre-existing CNV analysis, the CNV variants are included in all disease risk estimates in this work. In particular, in Figure 2, the data points for “full-exon sequencing” and “full-exon sequencing + special cases” already contain contributions from these CNV calls, while the “full-exon sequencing + special cases + CNV” adds in the additional panel-wide deletion CNVs that are not already performed as part of the Counsyl Family Prep Screen.

2 Census Weighting

To estimate the disease risks for a United States population, disease risks were first calculated on a per-ethnicity basis. Patient ethnicities were self-reported as one of the categories in Supplementary Table S1. The United States weighted disease risks were then estimated as a weighted average of the per-ethnicity values, with weights as given in Supplementary Table S1. These weights were obtained by mapping US census ethnicities onto the ethnicity categories reported on the Family Prep Screen test requisition form.

3 Disease Risk Calculation

The modeled fetal disease risk is the probability that a random conceptus is affected by at least one condition.

\[ R = P(\text{fetus affected by at least one disorder}) \]

We assume that the diseases are independent. Thus, we can expand the overall risk as a product over the per-disease risks \( R_d \):

\[ R = 1 - \prod_d (1 - R_d) = 1 - \prod_d (1 - P(\text{fetus affected by disease } d)) = 1 - \prod_d (1 - R_d) \]

3.1 Additive Approximation to Disease Risk

As written, the disease risk is unpleasant to deal with because it is non-additive in the disease components. However, by expanding the product to first order, we have a convenient additive approximation:

\[ R = 1 - \prod_d (1 - R_d) \approx \sum_d R_d \]

This approximation is justified because it facilitates interpretation by gene component and because the per-gene affected rates are all small for severe and profound diseases. We find that the error introduced by this approximation is on the order of \( 10^{-6} \) (relative error: \( 10^{-3} \)). For the current paper, we therefore use the additive disease risk for all calculations.
3.2 Example: Autosomal Recessive Disease Risk for one Disease

As an illustrative example, we consider the calculation of disease risk \( R_d \) for a single autosomal recessive condition. Here we make the simplifying assumption that the parental population could contain affected persons.

This illustration allows us to show that disease risk is a quadratic function of the allele frequencies with cross terms between different variants. This means that the disease risk has a snowball effect: the effect of a rare variant is amplified by the presence of other high-frequency variants.

In particular, suppose one has an existing targeted genotyping panel with two diseases and the ability to add exactly one additional variant to the panel. Suppose this additional variant can be chosen from either of the two diseases, and that the two candidate variants have the same AF. Then adding a variant in the higher-prevalence disease will contribute more DR than adding its counterpart from the lower-prevalence disease.

To begin our derivation, let’s suppose the disease in question is trivial and has a single pathogenic variant, with a per-chromosome pathogenic allele frequency of \( \pi \). The affected fetus will have two alleles, so the probability of being affected is

\[
P(\text{fetus has 2 pathogenic chromosomes}) = \pi^2
\]

Now let’s consider the case of a disease with exactly two loci with pathogenic alleles, each with frequency \( \pi_1, \pi_2 \). The probability of disease is then the probability that each fetus chromosome carries at least 1 pathogenic allele. The probability that one chromosome carries at least one pathogenic allele is

\[
P(\text{At least 1 pathogenic allele}) = \pi_1 + \pi_2 - \pi_1 \pi_2
\]

There are two chromosomes that are inherited independently, so we have

\[
P(\text{affected}) = (\pi_1 + \pi_2 - \pi_1 \pi_2)^2
\]

In the general case, we consider a disease with \( n \) independent pathogenic loci, in which case the DR is given by

\[
R_d = (1 - \prod_{i=1}^{n}(1-\pi_i))^2
\]

3.3 Calculating Disease Risk

The previous section described an example calculation of disease risk for autosomal recessive diseases in the case that affected individuals are able to reproduce and contribute to the pool of available chromosomes. However, the calculations quickly become intractable for diseases with complex inheritance and for the case that affected individuals are insufficiently healthy to reproduce. To deal with these different scenarios, we use a computer algebra system (Sympy) to calculate the disease risk, as was described previously [2].

3.4 Designing Optimal Targeted Genotyping Panels

Given a hypothetical targeted genotyping panel, how does one assess the importance of one or several variants? As discussed above, individual variants contribute non-additively (i.e., quadratically) to the overall disease risk. However, here we describe how, for certain applications, it is still possible to assign a disease risk number to each variant.

Suppose we have two sets of variants \( S_1 \) and \( S_2 \). We wish to assess the disease risk associated with the addition of the \( S_2 \) variants to an ECS panel containing the \( S_1 \) variants. We propose comparing the disease risks of original and final panel:

\[
R(S_1) \\
R(S_1 \cup S_2)
\]

Clearly, this metric depends on both sets of variants \( S_1 \) and \( S_2 \), so it does not give a unique metric for \( S_2 \). However, most of our applications will involve a single, well-defined order of variants—in which case the non-uniqueness is not an issue.

Based on this discussion, we are now in a place to describe our algorithm for designing optimal targeted genotyping panels. To determine the optimal rank-ordering of variants, let \( S = {} \) and apply the following algorithm:

1. For all \( v \not\in S \), calculate \( R_v = R(S \cup \{v\}) \)
2. Select \( v' = \text{argmax}(R_v) \)
3. Let \( S = S \cup \{v'\} \)

In practice, there are several possible simplifications to this algorithm. First, it is clear that one always selects the highest frequency variant within each disease. Thus, during a single iteration of the algorithm, only a single variant per disease need be considered. This reduces the number of risk calculations required during each iteration of the algorithm. Second, given the disease-additivity of the risk, one can pre-calculate the per-disease rankings and risk contributions.

4 References

[1] US census. [http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml](http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml). Accessed: 2016-1-26.

[2] I. S. Haque, G. A. Lazarin, H. P. Kang, E. A. Evans, J. D. Goldberg, and R. J. Wapner. Modeled fetal risk of genetic diseases identified by expanded carrier screening. *JAMA*, 316(7):734–742, 2016.

[3] M. S. Watson, G. R. Cutting, R. J. Desnick, D. A. Driscoll, K. Klinger, M. Mennuti, G. E. Palomaki, B. W. Popovich, V. M. Pratt, E. M. Rohlf, et al. Cystic fibrosis population carrier screening: 2004 revision of american college of medical genetics mutation panel. *Genetics in medicine*, 6(5):387–391, 2004.
Part II
Supplementary Tables and Figures
## Supplementary Table S1

| Code | Ethnicity                              | Patients | Census Weight (%) |
|------|----------------------------------------|----------|------------------|
| af   | African or African-American            | 29480    | 11.73            |
| aj   | Ashkenazi Jewish                       | 33886    | 1.83             |
| cj   | French Canadian or Cajun               | 1760     | 0.68             |
| co   | Mixed or Other Caucasian               | 114503   | 0.00             |
| ea   | East Asian                             | 23566    | 1.68             |
| fi   | Finnish                                | 247      | 0.21             |
| hi   | Hispanic                               | 39342    | 14.76            |
| me   | Middle Eastern                          | 8369     | 1.09             |
| na   | Native American                        | 797      | 0.65             |
| ne   | Northern European                      | 94652    | 55.58            |
| pi   | Pacific Islander                       | 990      | 0.27             |
| sa   | South Asian                            | 18256    | 0.96             |
| se   | Southeast Asian                        | 8062     | 1.53             |
| so   | Southern European                      | 14682    | 9.03             |
| uk   | Unknown                                | 86052    | 0.00             |

The number of patients and US census weights are given for each of the possible self-reported ethnicities on the Counsyl Family Prep Screen. The “uk” and “co” categories are given US weight 0 because they cannot be easily mapped to more specific ethnic categories.
6 Supplementary Figure S1

The curation workflow used to determine clinical significance of variants is summarized graphically. (a) The curation process is shown in the context of the overall laboratory workflow, in which inbound samples are eventually transformed into patient reports. (b) The curation workflow contributes lines of primary evidence that are reviewed manually, which are then combined with multiple lines of autogenerated supporting evidence to assess clinical significance.
## Supplementary Table S2

| Disease Name                                                                 | Gene      | Severity | Phenotypic Features: Tier 1                                      | Phenotypic Features: Tier 2                                      | Phenotypic Features: Tier 3                                      | OMIM      |
|------------------------------------------------------------------------------|-----------|----------|-----------------------------------------------------------------|-----------------------------------------------------------------|-----------------------------------------------------------------|-----------|
| Adrenoleukodystrophy                                                       | ABCD1     | Profound | intellectual disability shortened life span: infancy              | impaired mobility                                                | sensory impairment: vision                                       | #300100  |
| Mucopolysaccharidosis II (aka Hunter Syndrome)                              | IDS       | Profound | intellectual disability shortened life span: infancy              | impaired mobility                                                | sensory impairment: hearing dysmorphic features                  | #309900  |
| Peroxisome biogenesis disorder 4                                            | PEX6      | Profound | intellectual disability shortened life span: infancy              | impaired mobility                                                | sensory impairment: vision                                      | #614862  |
| Congenital disorder of deglycosylation                                      | NGLY1     | Profound | intellectual disability shortened life span: infancy              | impaired mobility                                                | dysmorphic features                                             | #615273  |
| Rhizomelic Chondrodysplasia Punctata Type 2                                  | GNPAT     | Profound | intellectual disability shortened life span: infancy              | internal physical malformation                                   | dysmorphic features immunodeficiency                             | #222765  |
| Dystrophinopathies (including Duchenne/Becker muscular dystrophy)           | DMD       | Severe   | intellectual disability                                          | impaired mobility shortened life span: premature adulthood       | internal physical malformation                                   | #310200  |
| Dyskeratosis congenita                                                       | RTEL1     | Severe   | intellectual disability shortened life span: childhood/adolescence| internal physical malformation                                   | dysmorphic features immunodeficiency                             | #615190  |
| Bardet-Biedl syndrome 9                                                      | BBS9      | Severe   | intellectual disability                                          | internal physical malformation                                   | sensory impairment: vision                                      | #615986  |
| Usher Syndrome Type 1D                                                       | CDH23     | Severe   | None                                                            | impaired mobility                                                | sensory impairment: vision                                      | #601067  |
| Dystrophic epidermolysis bullosa                                             | COL7A1    | Severe   | None                                                            | internal physical malformation                                   | sensory impairment: vision                                      | #226600  |
| Pseudoxanthoma Elasticum                                                     | ABCC6     | Moderate | None                                                            | internal physical malformation                                   | sensory impairment: vision                                      | #264800  |
| Persistent Mullerian Duct syndrome Type 1                                    | AMH       | Moderate | None                                                            | internal physical malformation                                   | immunodeficiency/cancer                                          | #261550  |
| Hermansky-Pudlak Syndrome                                                    | HP53      | Moderate | None                                                            | None                                                            | sensory impairment: vision                                      | #614072  |
| Lipoprotein lipase deficiency                                                | LPL       | Moderate | None                                                            | internal physical malformation                                   | sensory impairment: touch, other (including pain) dysmorphic features | #238600  |

Disease severity classification and phenotypic features are given for several diseases. These diseases were selected due to their presence on one or more existing carrier screening panels.
The disease risk attributed to panel-wide deletion calling is estimated as the difference in disease risk for panels with and without panel-wide deletion calling. Only genes for which any ethnicity has a panel-wide deletion disease risk $\geq 0.1$ per 100,000 are shown. Ethnicity abbreviations are enumerated in Supplementary Table S1.
Supplementary Figure S3

The percent contribution of panel-wide deletion calling is estimated as the relative difference in disease risk for panels with and without panel-wide deletion calling. The denominator is the disease risk in a panel with panel-wide deletion calling enabled.

| Gene   | US | af | aj | co | ea | hi | me | ne | sa | se | so | uk |
|--------|----|----|----|----|----|----|----|----|----|----|----|----|
| ACDAM  | 0  | 0  | 0  | 1  | 45 | 24 | 0  | 0  | 0  | 0  | 0  | 6  |
| ALDOB  | 4  | 0  | 5  | 1  | 6  | 0  | 3  | 0  | 0  | 7  | 0  |    |
| ATM    | 6  | 15 | 0  | 4  | 20 | 0  | 43 | 0  | 27 | 0  | 28 | 9  |
| ATP7B  | 1  | 40 | 0  | 2  | 4  | 0  | 0  | 0  | 0  | 0  | 0  |    |
| BBS1   | 5  | 0  | 0  | 0  | 45 | 0  | 3  | 0  | 0  | 45 | 0  |    |
| BLM    | 12 | 0  | 0  | 5  | 0  | 0  | 0  | 0  | 0  | 99 | 38 | 7  |
| CFTR   | 3  | 3  | 1  | 2  | 14 | 0  | 17 | 3  | 8  | 0  | 0  | 2  |
| DHC7R7 | 3  | 0  | 0  | 1  | 0  | 0  | 0  | 4  | 0  | 0  | 0  |    |
| DPYD   | 1  | 0  | 7  | 3  | 14 | 7  | 0  | 1  | 0  | 0  | 2  |    |
| FAH    | 10 | 0  | 0  | 0  | 0  | 0  | 7  | 0  | 0  | 28 | 0  |    |
| GAA    | 0  | 0  | 6  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  |    |
| GALT   | 6  | 0  | 96 | 2  | 0  | 0  | 0  | 5  | 0  | 0  | 16 |    |
| HBB    | 4  | 3  | 0  | 5  | 0  | 2  | 0  | 6  | 18 | 23 | 14 | 2  |
| PEX1   | 4  | 44 | 0  | 0  | 32 | 0  | 0  | 0  | 0  | 0  | 0  |    |
| PKHD1  | 3  | 0  | 10 | 4  | 0  | 6  | 0  | 3  | 0  | 0  | 0  | 7  |
| PMM2   | 3  | 0  | 3  | 2  | 0  | 0  | 0  | 2  | 0  | 0  | 9  | 0  |
| VPS13B | 24 | 32 | 9  | 25 | 28 | 61 | 78 | 19 | 0  | 69 | 36 | 34 |

Percent of DR contributed by deletion CNVs
10 Supplementary Table S3

| Gene | Variant |
|------|---------|
| CLN3 | NM_001042432.1(CLN3):c.461-280_677+382del966 |
| CTNS | NM_004937.2(CTNS):c.(?_36009)_(848?)del(aka 57 kb deletion) |
| GALC | NM_000153.3(GALC):c.1161+6532_polyA+9kb del(aka Ex11-17 del) |
| HEXA | NM_000520.4(HEXA):c.2564_253+5128delinsG(aka 7.6kb del) |
| MCOLN1 | NG_015806.1(MCOLN1):g.4127_10560del(aka 511_6944del) |
| NEB | NM_004543.4(NEB):c.(?_7431+1917)_(7536+373?)del |

_genes and variants for which specific CNV analysis is already performed as part of Family Prep Screen. No additional panel-wide CNV analysis was performed for these genes._
The per-variant contribution to modeled fetal disease risk (MFDR) is shown for each disease. Each tick represents the fraction of per-gene MFDR captured by an optimal, single-gene, ECS panel consisting of the $n$ most prevalent variants for that disease. Thus, the first tick for each disease represents the fractional MFDR of a single-variant panel for that disease; the second tick for each disease represents the fractional MFDR of a two-variant panel for that disease, and so on. The right sub-panel indicates the total MFDR of each disease. Diseases are displayed in decreasing order of disease risk. All MFDR values are US census weighted.
### 12 Supplementary Table S4

| GENE | NGS | TG | FOLD | | GENE | NGS | TG | FOLD |
|------|-----|----|------| |------|-----|----|------|
| HBB  | 42  | 39 | 1.1  | | LAMB3| 0.18| 0.045| 4.0  |
| CFTR | 33  | 27 | 1.2  | | SGCA | 0.18| 0.019| 9.2  |
| PAH  | 9.7 | 2.3| 4.3  | | FAH  | 0.17| 0.083| 2.1  |
| DHCR7| 6.6 | 5.5| 1.2  | | BLM  | 0.16| 0.027| 5.8  |
| PMM2 | 4.5 | 2.3| 2.0  | | ABCC8| 0.15| 0.069| 2.2  |
| DPYD | 4.2 | 1.7| 2.4  | | FANCC| 0.15| 0.061| 2.5  |
| ACADM| 4   | 2.9| 1.4  | | ASS1 | 0.15| 0.017| 8.6  |
| GAA  | 3   | 2.1| 1.5  | | NBN  | 0.13| 0.014| 9.3  |
| GBA  | 2.7 | 2.7| 1.0  | | SACS | 0.12| 0.00031| 380.9 |
| ALDOB| 2.7 | 2.1| 1.3  | | FANCC| 0.15| 0.017| 8.6  |
| ATP7B| 2.3 | 0.21| 11.2 | | PEX7 | 0.11| 0.074| 1.4  |
| PKHD1| 1.3 | 0.063| 19.9 | | DLD  | 0.1 | 0.046| 2.3  |
| ACADS| 1.2 | 0.48| 2.5  | | RMRP | 0.1 | 0.031| 3.4  |
| CTP2 | 1   | 0.68| 1.5  | | PPT1 | 0.1 | 0.045| 2.3  |
| GALT | 1   | 0.81| 1.3  | | SMPD1| 0.098| 0.032| 3.1  |
| CBS  | 0.87 | 0.015| 59.8 | | POMGNT1| 0.098| 0.0094| 10.5 |
| HEXA | 0.75 | 0.56| 1.3  | | BCKDHB| 0.092| 0.045| 2.1  |
| BBS1 | 0.71 | 0.52| 1.4  | | PCDH15| 0.075| 0.017| 4.5  |
| GALT | 0.47 | 0.12| 3.8  | | MAN2B1| 0.071| 0.0036| 20.1 |
| ACADVL| 0.44 | 0.2| 2.2  | | NAB | 0.066| 0.018| 3.6  |
| SLC26A2| 0.42 | 0.31| 1.4  | | CLRN1| 0.054| 0.038| 1.4  |
| ALPL | 0.41 | 0.068| 6.0  | | HSD17B4| 0.054| 0.013| 4.1  |
| BTD  | 0.4  | 0.14| 2.9  | | MCOLN1| 0.051| 0.041| 1.2  |
| ARSA | 0.37 | 0.19| 2.0  | | SLC37A4| 0.049| 0.021| 2.4  |
| ATM  | 0.37 | 0.071| 43.2 | | PROP1| 0.048| 0.011| 4.5  |
| AIRE | 0.35 | 0.02| 17.6 | | TMEM216| 0.046| 0.04| 1.1  |
| PEX1 | 0.33 | 0.14| 2.4  | | BCS1L| 0.04| 0.0028| 14.6 |
| CTNS | 0.33 | 0.18| 1.8  | | SLC17A5| 0.037| 0.0052| 7.1  |
| NPC1 | 0.31 | 0.016| 19.5 | | GRHPR| 0.031| 0.013| 2.4  |
| SLC22A5| 0.3 | 0.0022| 136.6 | | CLN5| 0.025| 0| inf |
| G6PC | 0.29 | 0.17| 1.7  | | MPI | 0.017| 0.0013| 12.9 |
| AGXT | 0.27 | 0.094| 2.9  | | AGA | 0.016| 4.5e-06| 3558.1 |
| IKBKAP| 0.26 | 0.26| 1.0  | | ALDH3A2| 0.015| 2.9e-05| 539.3 |
| VPS13B| 0.26 | 3.2e-06| 82582.0 | | LAMC2| 0.014| 2.2e-05| 640.7 |
| GCDH | 0.25 | 0.0085| 29.7 | | SLC12A6| 0.012| 7.3e-05| 163.0 |
| NPHS2| 0.25 | 0.051| 4.8  | | TTPA | 0.0099| 0.0038| 26.3 |
| BBS10| 0.25 | 0.065| 3.8  | | MLC1 | 0.0097| 0.00039| 25.0 |
| CLN3 | 0.23 | 0.2| 1.1  | | LAMA3| 0.0094| 6.2e-05| 151.0 |
| AGL  | 0.23 | 0.0027| 82.3 | | TH | 0.0077| 0.0015| 5.1  |
| ASPA | 0.21 | 0.17| 1.3  | | CPT1A| 0.0035| 3.3e-06| 1079.7 |
| NPHS1| 0.19 | 0.015| 13.1 | | OPA3 | 0.00048| 0.00025| 1.9  |
| TPP1 | 0.18 | 0.13| 1.4  | |     |     |     |     |

The disease risk per 100,000 births is enumerated for genes that were also present on the previous targeted genotyping version of Family Prep Screen. The MFDR is calculated for both the NGS and TG versions of the panel, along with the fold change of NGS over TG.
Pathogenic variants in *GBA* gene are enumerated in order of their contribution to the U.S. census-weighted disease risk for Gaucher Disease.

| variant name | pathogenic variants |
|--------------|---------------------|
| NM_001005741.2(GBA):c.1226A>G(N409S, aka N370S) | |
| NM_001005741.2(GBA):c.1448T>C(L483P, aka L444P) | |
| NM_001005741.2(GBA):c.1604G>A(R535H, aka R496H) | |
| NM_001005741.2(GBA):c.84dupG(aka p.L29Afs*18) | |
| NM_001005741.2(GBA):c.1342G>C(D448H) | |
| NM_001005741.2(GBA):c.115+1G>A(aka IVS2+1G>A) | |
| NM_001005741.2(GBA):c.1504C>T(R502C, aka R463C) | |
| NM_001005741.2(GBA):c.1297G>T(V433L, aka V394L) | |
| NM_001005741.2(GBA):c.1505G>A(R502H, aka R463H) | |
| NM_001005741.2(GBA):c.1343A>T(D448V, aka D409V) | |
The ACMG recommended 23, 29 and 42 variant CFTR TG panels are compared to NGS. Each bar represents the fraction of the total (NGS) per-ethnicity disease risk captured by different CFTR panels.