The Probability of Pathogenicity in Clinical Genetic Testing: A Solution for the Variant of Uncertain Significance

by (in alphabetical order)

Michael Gollob, Jeffrey S. Rosenthal, and Kevin Thorpe

Peter Munk Cardiac Centre and the University of Toronto

(November, 2015; last revised April 18, 2016)

1 Introduction

It frequently arises in medical genetics that a patient has a particular disease, and also has a genetic variant which may or may not be the cause of that disease. The variant is said to be of *unknown* or *uncertain significance* if its disease-causing probability cannot be determined, and this is a common challenge (see e.g. Richards et al. (2008), Cheon, Mozersky, and Cook-Deegan (2014), Domcheck and Weber (2008), and the references therein). The problem is put succinctly in the patient guide by Ambry Genetics (2015), which states, "Variants of unknown significance are DNA changes with too little information known to classify as either pathogenic or benign, and it is unknown whether they contribute to a medical condition."

Assessing whether or not the genetic variant did indeed cause the disease is important not only for future medical research and prevention, but also as a practical guide to whether or not the patient's family members are also at risk. However, in the case of a rare variant with little or no previous information available, it is unclear how this assessment should be made.

In this paper, we present a direct calculation for determining the probability that a rare genetic variant is indeed the cause of an observed disease. Our calculation requires one assumption, namely the natural-seeming "Variant Fraction Assumption" that in the absence of any other evidence, the probability that a newly observed rare genetic variant of a gene causes a specified disease is equal to the fraction of all previously-observed rare variants of that gene which did cause that disease (see Section 4.3 for details). With this one assumption, we are able to compute the desired probability purely in terms of the disease's prevalence in the population, and the fraction of rare variants in the general population and in the diseased population. Our results are described below.

2 Formal Set-Up

To set up our probability model, we make the following assumptions and notations:

- A certain disease D has a known prevalence p in the general population.
- Among the healthy population, a certain known fraction q have some distinct rare variant of a certain gene G.
- Among patients with the disease D, a certain fraction $r \geq q$ have some distinct rare variant of the gene G.
- Some subset S of the rare variants of G always cause the disease D, i.e. the probability of disease given a rare variant in S is 1.
- Rare variants of G which are *not* in S have no effect on D, i.e. the probability of disease given a rare variant *not* in S is the same as for patients without a rare variant.
- A new patient X is found to have a never-before-seen variant V of the gene G.

The question is, given all of the above, if X gets the disease D, then what is the probability that the genetic variant V was actually a *cause* of the disease D in X? That is, we wish to compute the conditional probability

$$\mathbf{P}(\mathrm{V} \in \mathrm{S} \mid \mathrm{X} \text{ has the disease D})\,,$$

i.e. the probability that X's variant V is in fact a *cause* of the disease D in X, given that X has the variant V and also has the disease D. We describe our calculation of this and related probabilities below.

We first note that since the genetic variant V has never been seen before, it is impossible compute its probabilities without making *some* additional assumption. Thus, we also make the *Variant Fraction Assumption* that in the absence of any other evidence, the probability that a newly observed genetic variant of a gene G is a cause of D, is equal to the fraction of *all* of the previously-observed rare variants of G which were indeed found to cause D. For a more precise statement of this assumption, see Section 4.3 below.

Remark. In fact, the assumption that $r \geq q$ follows directly from the other assumptions; see Appendix A2.

3 Main Result

Our main results are as follows. (Below, "unconditional probability" means the probability without conditioning on whether or not X has the disease D. Also, we write "V is the sole cause of D" to mean that V caused D, and furthermore patient X would not have gotten D in the hypothetical case that they instead did *not* have the rare variant V.)

Theorem 1 *For the above set-up, for a patient X having a rare variant V of gene G, under the Variant Fraction Assumption, we have the following probabilities, where* $y = (r-q)/(1-q)$*,* $z = p(1-y)/(1-py)$*,* $w = py \,/\, [pr + (1-p)q]$*, and* $u = (1-z)w$*.*
(a) The unconditional probability that the variant V is a cause of the disease D is given by:

$$\mathbf{P}(V \in S) \;=\; w\,.$$

(b) The unconditional probability that patient X will get disease D is given by:

$$\mathbf{P}(X \text{ gets } D) \;=\; z + u\,.$$

(c) Conditional on patient X getting the disease D, the conditional probability that the variant V was the sole *cause of D is given by:*

$$\mathbf{P}(\text{X's disease D was caused solely by V} \mid \text{X has D}) \;=\; \frac{u}{z+u}\,.$$

(d) Conditional on patient X getting the disease D, the conditional probability that the variant V is a cause of D is given by:

$$\mathbf{P}(V \in S \mid X \text{ gets } D) \;=\; \frac{w}{z+u} \;=\; \frac{r-q}{r(1-q)}\,.$$

Theorem 1 thus gives precise probabilities for the relevant possibilities related to patient X and disease D and rare genetic variant V. In particular, Theorem 1(d) gives a precise estimate of the probability, given that patient X has disease D and has the rare genetic variant V, that V is in fact a cause of the disease D.

Theorem 1 is proved in the next section. We first consider some numerical examples.

For example, if $p = 1/4,000$ is the prevalence of the disease in the general population, and $q = 2\% = 0.02$ is prevalence of *some* rare variant of G in the healthy population, and $r = 40\% = 0.4$ is the prevalence of *some* rare variant of G in patients with the disease, then conditional on X getting the disease, the probability that the variant V is a cause of the disease works out to: $\mathbf{P}(V \in S \mid X \text{ gets } D) = 0.9699815 \doteq 97.0\%$, or about 97 percent (i.e., nearly certain).

Or, if $p = 1/400$ and $q = 0.01$ and $r = 0.4$, then $\mathbf{P}(V \in S \mid X \text{ gets } D) \doteq 98.5\%$.

By contrast, if $p = 1/50$ and $q = 0.1$ and $r = 0.2$, then $\mathbf{P}(V \in S \mid X \text{ gets } D) \doteq 58.1\%$, which is much smaller.

Or, if $p = 1/400$ and $q = 0.1$ and $r = 0.15$, then $\mathbf{P}(V \in S \mid X \text{ gets } D) \doteq 39.5\%$.

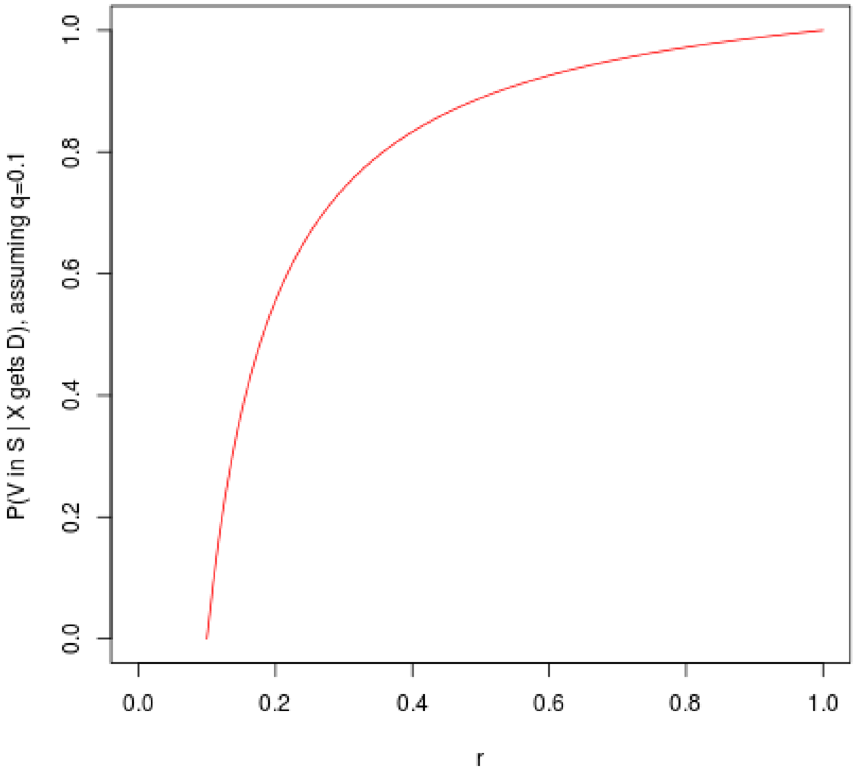

Figure 1: Disease cause probabilities as a function of r, with $q = 0.1$ fixed.

A plot of other values of $\mathbf{P}(V \in S \mid X \text{ gets } D)$ is presented in Figure 1 as a function of r, with $q = 0.1$ fixed.

Probabilities for other parameter values can be computed using the above formulae, or using our simple javascript online calculator available at: `www.probability.ca/pathprob`

As a further check, we note that the formula in Theorem 1(d) gives answers which make sense even for certain extreme parameter values. For example, if $r = 1$ (i.e., 100%), then it is computed that the formula gives a value of 1. This makes sense, since if $r = 1$, then by the definition of r, *every* diseased patient has a rare variant of G, which means that the disease can *only* be caused by a rare variant of G, so that V must indeed cause D.

Or, if $q = 0$, then again it is computed that the formula gives a value of 1. This also makes sense, since if $q = 0$, then by the definition of q, *no* healthy patients have rare variants of G, which means that rare variants of G *always* cause D, so again V must indeed cause D.

By contrast, if $r = q$, then it is computed that the formula gives a value of 0. This again makes sense, since if the rate of rare variants of G is the same for diseased and healthy patients, then the rare variants of G do *not* cause any additional disease at all, so none of them cause D.

As a final comment, we note that since $u = (1-z)w$, the formula in Theorem 1(c) is smaller than that in Theorem 1(d) by a factor of $(1-z)$. This is due to the possibility that $V \in S$ but X still "would" have gotten D by chance alone, i.e. that V does indeed cause D, but X would have gotten D even in the absence of V (e.g. from a variant of some other gene besides G). Now, usually z will be very small, so the answers in (c) and (d) will be very similar, though not identical.

4 Proof of Theorem 1

In this section, we prove Theorem 1 using a sequence of probability calculations. We break up our argument into several steps.

4.1 Preliminary Population Prevalence Calculations

We first note that our assumption above that rare variants of G which are *not* in S have no effect on D, can be written more formally as

$$\mathbf{P}(\text{X has D} \mid \text{X has a rare variant V} \notin \text{S}) \;=\; \mathbf{P}(\text{X has D} \mid \text{X has no rare variant})\,.$$

From this it follows (see Appendix A1) that

$$\begin{aligned}&\mathbf{P}(\text{X has rare variant V} \notin \text{S} \mid \text{X has D, and no variant in S})\\ &\qquad = \mathbf{P}(\text{X has some rare variant} \mid \text{X does not have D})\,.\end{aligned}$$

We next calculate two population fractions that will be important in our solution.

First, we write y for the fraction of diseased patients who have a rare variant of G which is in S, i.e. which does cause D. (Thus, y is close to r, but slightly less since even among diseased patients without a variant in S, a fraction q of them will still happen to have a rare variant not in S by chance alone, just like for the healthy population.)

In terms of y, the set of all diseased patients can be divided into three groups: a fraction y with a rare variant of G which is in S, a fraction $(1-y)q$ with a rare variant of G which is *not* in S, and a fraction $(1-y)(1-q)$ with no rare variant of G at all. Hence, the fraction of diseased patients with *some* rare variant of G is equal to $y + (1-y)q$.

On the other hand, we know that the overall fraction of diseased patients with some rare variant of G is equal to r. For this to hold, we must have $y + (1-y)q = r$. Solving for y, we obtain that $y = (r-q)/(1-q)$.

Then, since a fraction p of the population is diseased, and a fraction y of diseased patients have a variant in S, it follows that the fraction of the total population who have a rare variant in S is equal to the product py. And, the fraction who do *not* have a rare variant in S is equal to $1 - py$.

Next, we write z for the prevalence of the disease D among all people who specifically do *not* have a variant in S. Then the fraction of people who have D but do *not* have a variant in S is equal to $(1 - py)z$. And, the fraction of people with a variant in S (who therefore have D) is equal to py. So, the total fraction of the population who have the disease D is equal to $(1 - py)z + py$.

On the other hand, we know that the overall prevalence of the disease is equal to p. For this to hold, we must have $p = py + (1 - py)z$. Solving for z, we compute that $z = p(1 - y)/(1 - py)$.

4.2 The Prior Probability of Disease

Prior to diagnosis, what was the prior probability that a given patient X, who has some rare variant V, would get the disease D? That is, what is the conditional probability that X gets D, conditioning (throughout) on the fact that X has a rare variant V?

To answer this question, let I be the indicator function of the subset S, so I(V)=1 if V $\in$ S, otherwise I(V)=0. We know that the prior probability of X getting the disease D is equal to 1 (i.e., 100%) if I(V)=1, or is equal to z if I(V)=0. So, if we *knew* I(V), i.e. if we *knew* whether or not V causes D, then could write this as:

$$\mathbf{P}(\text{X gets D} \,|\, \text{I(V)}) \;=\; z \,+\, (1 - z) \times I(V)\,.$$

In fact we do not know I(V), i.e. it could equal either 1 or 0. So, instead, we use the Law of Total Expectation (that the expected value of a conditional probability is the unconditional probability). This shows that:

$$\mathbf{P}(X \text{ gets } D) \;=\; \mathbf{E}\Big[\mathbf{P}(\text{X gets D} \,|\, \text{I(V)})\Big] \;=\; z \,+\, (1 - z) \times \mathbf{E}[I(V)]$$

$$=\; z \,+\, (1 - z) \times \mathbf{P}[I(V) = 1] \;=\; z \,+\, (1 - z) \times \mathbf{P}(V \in S)\,.$$

This gives a formula for the prior probability that X would get D, in terms of the probability $\mathbf{P}(V \in S)$ that V is in S. However, we do not know $\mathbf{P}(V \in S)$, so it must be estimated.

4.3 The Variant Fraction Assumption

To continue, we need to obtain an estimate for $\mathbf{P}(V \in S)$, the unconditional probability (in the absence of any other evidence) that the newly observed variant V of G is in fact a cause of the disease D. Now, since V was never before seen, there is no way to directly calculate this probability. Instead, as mentioned above, we use the *Variant Fraction Assumption* that in the absence of any other evidence, the probability that a newly observed genetic variant of a gene G is a cause of D, is equal to the fraction of *all* of the previously-observed rare variants of G which were indeed found to cause D. That is, we assume that

$$\mathbf{P}(V \in S) \;=\; \frac{\text{fraction of the population with a disease-causing rare variant of G}}{\text{fraction of the population with any rare variant of G}}\,.$$

This assumption appears to be quite reasonable, in the absence of any other prior information about the new variant V. In any case, some such assumption must be made, otherwise no probabilities associated with V can possibly be computed. But under this one assumption, all of the remaining probability calculations can be completed.

4.4 Estimating the Variant Probability

Using the above Variant Fraction Assumption, we are able to compute the desired probability $\mathbf{P}(V \in S)$. To do this, we need to compute the fraction of *all* of G's rare variants which are in S. We proceed as follows.

Since a fraction p of the population is diseased, and since a fraction r of them have some rare variant of G, it follows that the fraction of the population who are diseased *and* have some rare variant of G is pr. Similarly, since y is the fraction of diseased patients who have a variant in S, it follows that the fraction of the population who are diseased and have a rare variant in S is py. Also, the fraction of the population who are *healthy* and have *some* rare variant of G is $(1-p)q$.

That is, a fraction $pr + (1-p)q$ of the population has a rare variant, and a fraction py of the population has a rare variant in S. So, assuming these rare variants are all *distinct*, the fraction of all the rare variants of G in the entire population which are in S is equal to $py/[pr + (1-p)q]$.

We then estimate the probability $\mathbf{P}(V \in S)$ by the above fraction of all the rare variants of G which are in S, i.e. by

$$\mathbf{P}(V \in S) \;\approx\; py \,/\, [pr + (1-p)q] \;=:\, w$$

where $w = py \,/\, [pr + (1-p)q]$, as claimed in Theorem 1(a).

Now, since we earlier derived the prior-to-diagnosis probability $\mathbf{P}(\text{X gets D}) = z + (1 - z) \times \mathbf{P}(V \in S)$, it then follows from the above estimate that

$$\mathbf{P}(X \text{ gets } D) \;=\; z \,+\, (1-z) \times w \;=:\; z \,+\, u$$

where $u = (1-z)w$, as claimed in Theorem 1(b).

4.5 The Cause Probability

The above formula for **P**(X gets D) can be interpreted as follows: Patient X can get the disease D either without any influence at all from the gene G (with probability z), or caused by the variant V of G (with probability f).

Under this interpretation, given that X *does* in fact have the disease D, the conditional probability that the disease D in X was *caused* by the genetic variant V, and *would not otherwise have arisen*, is given by the second probability divided by the sum of the two probabilities, i.e. by:

$$\mathbf{P}(\text{X's disease D was caused solely by V} \mid \text{X has D}) \;=\; \frac{u}{z+u}\,,$$

as claimed in Theorem 1(c).

This formula thus gives an estimate of the probability, given that patient X has disease D, that the disease was caused solely by their rare genetic variant V of the gene G. This is similar to, but not quite the same as, our desired conditional probability, as we now explain.

4.6 Computing the Conditional Probability

Putting the previous equations together, we can compute the required conditional probability, as follows:

$$\begin{aligned}\mathbf{P}(V\in S \mid X \text{ gets } D) \;&=\; \frac{\mathbf{P}(V\in S,\text{ and } X \text{ gets } D)}{\mathbf{P}(X \text{ gets } D)} \\ &=\; \frac{\mathbf{P}(V\in S)}{\mathbf{P}(X \text{ gets } D)} \;=\; \frac{w}{z+u}\,.\end{aligned}$$

This gives our first formula claimed in Theorem 1(d).

Finally, through careful algebraic simplification, it is verified directly that in fact $\frac{w}{z+u} = \frac{r-q}{r(1-q)}$ (which, surprisingly, does not depend on p), thus giving the second formula claimed in Theorem 1(d).

Remark. The above equations can be combined as follows. We have that

$$\begin{aligned}\frac{u}{z+u} \;&=\; \mathbf{P}(\text{D was caused solely by V} \mid \text{X gets D}) \\ &=\; \text{P(D was caused solely by V) / P(X gets D)} \\ &=\; \mathbf{P}(\text{D was caused solely by V} \mid \text{V in S})\, \mathbf{P}(\text{V in S} \mid \text{X gets D}) \\ &=\; \mathbf{P}(\text{D was caused solely by V} \mid \text{V in S})\, \frac{w}{z+u}\,,\end{aligned}$$

whence

$$\mathbf{P}(\text{D was caused solely by V} \mid \text{V in S}) \;=\; \frac{u}{w} \;=\; \frac{(1-z)w}{w} \;=\; 1-z\,.$$

Taking the complementary event,

$$\mathbf{P}(\text{X would have gotten D even without having V} \mid \text{V in S}) \;=\; z\,,$$

which makes sense since z is the prevalence of the disease in people who do *not* have a rare variant, corresponding to the appropriate hypothetical.

Final Remark. After originally completing this research, we became aware of the related recent paper by Ruklisa, Ware, Walsh, Balding, and Cook (2015). That paper mostly focuses on specific probability estimates for specific diseases and specific genetic variants. However, it does begin with what they call the "prior odds of pathogenicity", which appears to correspond to odds for the event $\mathbf{P}(V \in S \mid X \text{ gets } D)$ considered in Theorem 1(d) above. For that case, they assert that the odds "might be assumed to be" given by

$$\frac{\text{Burden of rare variants in cases} - \text{Burden of rare variants in controls}}{\text{Burden of rare variants in controls}}\,.$$

In our notation, this appears to correspond to asserting that

$$\frac{\mathbf{P}(V \in S \mid X \text{ gets } D)}{1 - \mathbf{P}(V \in S \mid X \text{ gets } D)} \;=\; \frac{r-q}{q}\,,$$

or equivalently that

$$\mathbf{P}(V \in S \mid X \text{ gets } D) \;=\; \frac{\frac{r-q}{q}}{1 + \frac{r-q}{q}} \;=\; \frac{r-q}{r}\,.$$

This last expression is fairly similar to the final formula in our Theorem 1(d), but it differs by a factor of $1-q$. It appears that the reason for this discrepency is their assertion that "it is reasonable to assume that the burden of benign rare variants in cases is equal to the burden of rare variants in controls", which apparently corresponds to assuming that

$$\mathbf{P}(\text{Rare variant not in S} \mid \text{Diseased}) \;=\; \mathbf{P}(\text{Rare variant not in S} \mid \text{Not diseased})\,,$$

which is different from what we believe (for more see Appendix A3).

Appendix: Additional Probability Calculations

We here provide a few additional probability calculations, to further clarify some of the material in the main text. For these calculations, define a through f to be the proportions of the total population in each of the six categories implied by the following Table:

	Diseased	Not diseased
Rare variant in S	a	d
Rare variant not in S	b	e
No rare variant	c	f

A1. We first show that the assumption

$$\mathbf{P}(\text{X has D} \mid \text{X has a rare variant V} \notin \text{S}) \;=\; \mathbf{P}(\text{X has D} \mid \text{X has no rare variant}) \qquad (*)\,.$$

implies the condition that

$$\begin{aligned}&\mathbf{P}(\text{X has rare variant V} \notin \text{S} \mid \text{X has D, and no variant in S})\\ &\qquad = \mathbf{P}(\text{X has some rare variant} \mid \text{X does not have D})\,. \qquad (**)\end{aligned}$$

In terms of the above Table, condition $(*)$ is equivalent to saying that $b/(b+e) = c/(c+f)$, i.e. that $b/e = c/f$. Meanwhile, condition $(**)$ is equivalent to saying that $b/(b+c) = (d+e)/(d+e+f)$. On the other hand, our assumption that variants in S always cause the disease D implies that $d = 0$. Hence, if $b/e = c/f$, then $(d+e)/(d+e+f) = e/(e+f) = b/(b+c)$, thus establishing $(**)$.

A2. We here show that the assumption $r \geq q$ actually follows from our other assumptions. Indeed, in terms of the above Table, using as above that $d = 0$ and $e/(e+f) = b/(b+c)$, we have that

$$q \;=\; \mathbf{P}(\text{Rare variant} \mid \text{Not Diseased}) \;=\; \frac{d+e}{d+e+f} \;=\; \frac{e}{e+f} \;=\; \frac{b}{b+c}\,.$$

Also

$$r \;=\; \mathbf{P}(\text{Rare variant} \mid \text{Diseased}) \;=\; \frac{a+b}{a+b+c}\,.$$

The result then follows since

$$r - q \;=\; \frac{a+b}{a+b+c} - \frac{b}{b+c} \;=\; \frac{(ab+ac+b^2+bc)-(ab+b^2+bc)}{(a+b+c)(b+c)} \;=\; \frac{ac}{(a+b+c)(b+c)} \;\geq\; 0\,.$$

A3. Finally, we consider the quantities which arise in the reference Ruklisa et al. (2015), as discussed in our Final Remark above, namely

$$\mathbf{P}(\text{Rare variant not in S} \mid \text{Diseased}) \qquad (\&)$$

and

$$\mathbf{P}(\text{Rare variant not in S} \mid \text{Not diseased})\,. \qquad (\&\&)$$

In our notation, $(\&\&) = \mathbf{P}(\text{Rare variant not in S} \mid \text{Not diseased}) = q$ which, in terms of the above Table, is equal to $e/(e+f)$. By contrast, $(\&) = \mathbf{P}(\text{Rare variant not in S} \mid \text{Diseased}) = b/(a+b+c)$. Hence, assuming $(*)$, we have $(\&) = b/(a+b+c) \leq b/(b+c) = e/(e+f) = (\&\&)$. Furthermore, we have non-zero equality only when $a = 0$, which corresponds to no variants in S, i.e. to the gene G having no effect whatsoever on the disease D. Otherwise, $a > 0$, and hence $(\&) < (\&\&)$ (assuming $b > 0$), leading to a different result from that of Ruklisa et al.

References

[1] Ambry Genetics (2015), Your genetic test result. Available at: http://www.ambrygen.com/sites/default/files/ General%20VUS%20patient%20brochure%20Updates_0.pdf

[2] Cheon, J.Y., Mozersky, J., and Cook-Deegan, R. (2014), Variants of uncertain significance in BRCA: a harbinger of ethical and policy issues to come? *Genome Medicine* **6:121**, 1–10.

[3] Domchek, S. and Weber, B.L. (2008), Genetic variants of uncertain significance: flies in the ointment. *Journal of Clinical Oncology* **26(1)**, 16–17.

[4] Richards, C.S., Bale, S., Bellissimo, D.B., Das, S., Grody, W.W., Hegde, M.R., Lyon, E., and Ward, B.E. (2008), ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genetics in Medicine* bf 10(4), 294–300.

[5] Ruklisa, D., Ware, J.S., Walsh, R., Balding, D.J., and Cook, S.A. (2015), Bayesian models for syndrome- and gene-specific probabilities of novel variant pathogenicity. *Genome Medicine* **7:5**, 1–16.