## 1 One-Sided Matching Portal (OSMP): a tool to facilitate rare disease patient matchmaking

## 2 Authors:

- 3 Matthew Osmond,<sup>1†</sup> E. Magda Price,<sup>1†</sup> Orion J. Buske,<sup>2</sup> Mackenzie Frew,<sup>3</sup> Madeline Couse,<sup>3</sup> Taila
- 4 Hartley,<sup>1</sup> Conor Klamann,<sup>3</sup> Hannah G. B. H. Le,<sup>3</sup> Jenny Xu,<sup>3</sup> Delvin So,<sup>3</sup> Anjali Jain,<sup>3</sup> Kevin Lu,<sup>3</sup> Kevin Mo,<sup>3</sup>
- 5 Hannah Wyllie,<sup>1</sup> Erika Wall,<sup>1</sup> Hannah G. Driver,<sup>1</sup> Warren A. Cheung,<sup>4</sup> Ana S.A. Cohen,<sup>4,5,6</sup> Emily G.
- 6 Farrow,<sup>4,5,6</sup> Isabelle Thiffault,<sup>4,5,6</sup> Care4Rare Canada Consortium,<sup>1</sup> Andrei L. Turinsky,<sup>3</sup> Tomi Pastinen,<sup>4,5,7</sup>
- 7 Michael Brudno,<sup>1,2,3,8,9,10</sup> Kym M. Boycott<sup>1</sup>

## 8 Affiliations:

- 9 <sup>1</sup> Children's Hospital of Eastern Ontario Research Institute, University of Ottawa, Ottawa, Canada
- 10 <sup>2</sup> PhenoTips, Toronto, Canada
- <sup>3</sup> Centre for Computational Medicine, The Hospital for Sick Children, Toronto, Canada
- 12 <sup>4</sup> Genomic Medicine Center, Children's Mercy Hospital, Kansas City, Missouri, USA
- <sup>5</sup> University of Missouri Kansas City School of Medicine, Kansas City, Missouri, USA
- <sup>6</sup> Department of Pathology and Laboratory Medicine, Children's Mercy Hospital, Kansas City, Missouri,
- 15 USA
- 16 <sup>7</sup> Children's Mercy Research Institute, Kansas City, Missouri, USA
- 17 <sup>8</sup> UHN DATA Team, University Health Network, Toronto, Canada
- <sup>9</sup> Department of Computer Science, University of Toronto, Toronto, Canada
- 19 <sup>10</sup> Vector Institute, Toronto, Canada
- 20 <sup>+</sup> These two co-authors contributed equally to this work
- 21 **Correspondence to:** Kym M Boycott, Department of Genetics, CHEO, 401 Smyth Rd, Ottawa, ON, K1H
- 22 8L1, Canada; <u>kboycott@cheo.on.ca</u> and Matthew Osmond, Research Coordinator, CHEO, 401 Smyth Rd,
- 23 Ottawa, ON, K1H 8L1, Canada; MaOsmond@cheo.on.ca

# 24 Abstract (max 350 words)

25	Background: Genomic matchmaking - the process of identifying multiple individuals with overlapping
26	phenotypes and rare variants in the same gene - is an important tool facilitating gene discoveries for
27	unsolved rare genetic disease (RGD) patients. Current approaches are two-sided, meaning both patients
28	being matched must have the same candidate gene flagged. This limits the number of unsolved RGD
29	patients eligible for matchmaking. A one-sided approach to matchmaking, in which a gene of interest is
30	queried directly in the genome-wide sequencing data of RGD patients, would make matchmaking
31	possible for previously undiscoverable individuals. However, platforms and workflows for this approach
32	have not been well established.
33	Results: We released a beta version of the One-Sided Matching Portal (OSMP), a platform capable of
34	performing one-sided matchmaking queries across thousands of participants stored in genomic
35	databases. The OSMP returns variant-level and participant-level information on each variant occurrence
36	(VO) identified in a queried gene and displays this information through a customizable data table. A
37	workflow for one-sided matchmaking was developed so that researchers could effectively prioritize the
38	many VOs returned from a given query. This workflow was then tested through pilot studies where two
39	sets of genes were queried in over 2,500 individuals: 130 genes that were newly associated with disease
40	in OMIM, and 178 candidate genes that were not yet associated with a described disease-gene
41	association in OMIM. These pilots both returned a large number of initial VOs (12,872 and 20,308,
42	respectively), however the workflow successfully filtered out over 99.8% of these VOs before they were
43	sent for review by a patient's clinician. Filters on participant-level information, such as variant zygosity,
44	participant phenotype, and whether a variant was also present in unaffected participants were
45	especially effective in this workflow at reducing the number of false positive matches.

- 46 **Conclusions:** As demonstrated through the two pilot studies, one-sided matchmaking queries can be
- 47 efficiently performed using the OSMP. The availability of variant-level and participant-level data is key to
- 48 ensuring this approach is practical for researchers. In the future, the OSMP will be connected to
- 49 additional RD databases to increase the accessibility of matchmaking to unsolved RGD patients.

- 51 **Keywords:** Rare disease, rare genetic disease, matchmaking, data sharing, genome wide sequencing,
- 52 OMIM, disease gene discovery, one-sided matchmaking

## 53 Background

54	In the past decade, genome-wide sequencing (GWS), including both exome sequencing and genome
55	sequencing, has become a standard tool for the diagnosis of rare genetic diseases (RGDs) in many
56	regions of the world. Depending on the indication(s) and specific implementation, it is estimated that
57	approximately 40% of RGD families tested by GWS will receive a diagnosis. <sup>1</sup> For comparison, the first-tier
58	chromosomal microarray (CMA) test has an average diagnostic yield of 12.2% across a range of
59	indications including developmental delay, dysmorphic features, intellectual disability, learning
60	disabilities, autism and multiple congenital anomalies. <sup>2</sup> And yet despite the success of clinical GWS
61	testing, a significant portion of families with RGD remain undiagnosed following their clinical testing.
62	Families without a diagnosis following clinical GWS testing fall broadly into three categories: 1) the
63	etiology of their condition is not monogenic, and unlikely to be resolved by current genetic testing
64	methods; 2) the genetic etiology of their condition is not detectable by the GWS method employed, e.g.
65	a complex structural variant with exome sequencing, but might be resolved by another technology; and
66	3) the genetic etiology of their condition is detectable by GWS, but the gene, region, or variant cannot
67	yet be associated with disease, e.g., due to analytical challenges or insufficient evidence. In other words,
68	families in this final group have a diagnosis within their existing sequencing data, but it is not recognized
69	at the time of clinical testing.

Given the rapid pace at which new disease-gene associations and disease-variant associations are
discovered each year, there is great interest in revisiting clinically generated GWS data to search for
undetected molecular causes of RGDs. Many large-scale RGD research programs like the Undiagnosed
Disease Network (USA, https://undiagnosed.hms.harvard.edu/), Genomics Answers for Kids (USA,
https://www.childrensmercy.org/childrens-mercy-research-institute/studies-and-trials/genomicanswers-for-kids/), RD-Connect (EU, https://rd-connect.eu/) and our program discussed here, Care4Rare

(Canada, https://www.care4rare.ca/), leverage this approach – conducting "reanalysis" of clinically 76 77 generated GWS data to search for undetected molecular causes of RGD. A review of publications 78 involving reanalysis of clinical GWS found the median new diagnostic rate was 15%, but that it varied 79 considerably between studies.<sup>3</sup> Working with GWS data in a research context broadens the tools and 80 approaches that can be used to identify the molecular cause of the RGD. We recently performed a 81 "clinical reanalysis" of a cohort of 287 families undiagnosed following clinical GWS.<sup>4</sup> The reanalysis, 82 limited to variants in genes known to be associated with disease in the Online Mendelian Inheritance in 83 Man (OMIM) database, led to the identification of compelling candidate variants in 39 families (14%), 84 and ultimately resulted in diagnoses for 13 families (5%). The most common factor in making these new 85 diagnoses was the availability of new genomic knowledge, including new disease-gene associations, 86 disease-variant associations, and phenotype expansions of existing conditions. These findings indicate 87 that some undiagnosed RGD families will receive a diagnosis due to advances in global genomics 88 knowledge if their data is revisited over time.

89 While periodic clinical reanalysis may capture diagnoses in newly described disease associations, it fails 90 to address facilitating the discovery of the estimated thousands of novel disease-genes in real-time.<sup>5</sup> 91 RGD families with conditions with an unknown molecular etiology will remain undiagnosed until enough 92 evidence is generated to associate a causal gene with their specific disease. An important criterion to 93 substantiate such an association is the identification of multiple unrelated probands with pathogenic 94 variants in the same gene and overlapping phenotype such that they are believed to have the same 95 undescribed disease.<sup>6</sup> The process of identifying similar families in this way is broadly called genomic 96 matchmaking. Since RGDs are, by definition, extremely rare, the discovery of new disease genes hinges 97 on worldwide sharing of information about undiagnosed families. To do this, tools like the Matchmaker 98 Exchange (MME) enable submitters (who might be researchers, clinicians or RGD families) from around 99 the world to connect and discuss potential overlap of cases.<sup>7</sup> The MME uses a two-sided approach to

100 matchmaking, in which submitters will only be matched if they have the same candidate gene flagged 101 for their respective families. Since its inception in 2015, the MME has been heavily used by both 102 research and clinical rare disease communities to facilitate the discovery of hundreds of novel disease-103 gene associations.<sup>7</sup> Care4Rare's own experience with two-sided matchmaking through the MME reflect 104 its utility; over just two years, Care4Rare matched on 194 novel candidate genes, resulting in 861 105 connections with other submitters, ultimately leading to collaborations for 23 (15%) of these genes.<sup>8</sup> 106 While two-sided matchmaking has been a fundamental tool for gene discovery, the accessibility of this 107 approach is limited for several reasons. Firstly, this hypothesis-based matching requires that each family 108 have a candidate gene flagged (and thus must have been analyzed/reviewed) before it can be 109 submitted. Secondly, while most databases currently connected to the MME support the inclusion of 110 participant-level information (e.g., specific variant, zygosity, detailed phenotype, inheritance) when 111 submitting a candidate gene for matchmaking, in our experience, most matches lack this information at 112 the outset. Additional details on potential matches therefore must be exchanged by email after the 113 initial match is made, which is time consuming for users.<sup>8</sup> The current two-sided matchmaking model 114 thus limits sharable data to the small set of families who: i) have been (re)analyzed and ii) have a flagged 115 candidate gene, and limits submitters to those with the resources to follow up on many potential 116 matches by email. Given the amount of GWS data that is produced through clinical and research 117 testing,<sup>9,10</sup> there are undoubtedly many untapped RGD datasets not currently available for two-sided 118 matchmaking, restricting the ability to identify novel genetic etiologies for undiagnosed families. 119 Additional approaches to genomic matchmaking, which aim to address some of the shortcomings 120 identified in two-sided matchmaking, have been proposed.<sup>7</sup> One-sided matchmaking, involves a single 121 party submitting a query on a gene or variant of interest to a database of GWS data to identify 122 undiagnosed participants with variants matching the initial query. While a network of RGD databases 123 like the MME has not yet been established for one-sided matchmaking, multiple databases have

designed their own approaches to this type of matchmaking. MyGene2

125 (https://mygene2.org/MyGene2/), Geno2MP (https://geno2mp.gs.washington.edu/), VariantMatcher 126 (https://variantmatcher.org/), and Franklin (https://franklin.genoox.com/) have designed variant-level 127 implementations of one-sided matchmaking, in which users can search for the presence of a specific 128 variant within the database, and receive phenotypic information on any participant found to carry this 129 variant.<sup>11</sup> Other platforms, such as the DatabasE of genomiC Variation and Phenotype in Humans using Ensembl Resources (DECIPHER),<sup>12</sup> RD-Connect Genome-Phenome Analysis Platform (GPAP),<sup>13</sup> and sear<sup>14</sup> 130 131 support a gene-level approach to one-sided matchmaking instead, where a user can query a gene of 132 interest and all variants in this gene are returned by the queried database. In evaluating these existing 133 approaches to one-sided matchmaking, we see gene-level one-sided matchmaking as having the 134 potential to identify variants of interest in undiagnosed RGD families through the querying of two types 135 of genes: 1) genes that have been recently associated with human disease; and 2) genes that have not 136 yet been associated with a disease. Given that a gene-level approach to one-sided matchmaking has the 137 potential to, depending on the queried database's size, return many variants, we expect additional data 138 related to these variants will be crucial in ruling out false positive matches. To our knowledge, however, 139 there has not been an assessment of what types of data and level of detail are needed to allow one-140 sided matchmaking users to filter gene-level queries down to a manageable number of variants of 141 potential interest.

In this paper, we present a workflow for gene-level one-sided matchmaking and a beta version of a tool to support this approach, called the One-Sided Matching Portal (OSMP). The platform was designed with an emphasis towards providing a variety of participant-level and variant-level information inside a customizable interface, to best support users in making efficient one-sided matchmaking queries and limit the external communications required to rule potential matches in or out. To test the utility of the OSMP, as well as the one-sided matchmaking approach in general, we ran pilot studies using two sets of

- 148 genes to identify new variants of interest in participants enrolled in Care4Rare: 1) 130 newly described
- 149 OMIM disease genes to search for disease-causing variants that would not have been recognized at the
- time of analysis by the clinical diagnostic laboratory; and 2) 178 novel candidate genes previously
- 151 flagged on GWS analysis performed by the Care4Rare program.
- 152 Methods
- 153 The One-Sided Matchmaking Portal (OSMP)
- 154 We designed and built a tool, called the One-Sided Matching Portal (OSMP,
- 155 <u>https://github.com/ccmbioinfo/osmp</u>), a web-based portal that can be connected to one or more RGD
- databases that contain variant and health information from research participants. The beta version of
- 157 the OSMP supports gene-based queries of PhenoTips<sup>®</sup> instances by fetching and displaying single
- 158 nucleotide variants and small insertions or deletions from the PhenoTips<sup>®</sup> variant store and participant
- 159 information from PhenoTips<sup>®</sup> participant records.<sup>15</sup> The OSMP's frontend is written using the React
- 160 JavaScript library (<u>https://react.dev/</u>), and the backend is designed using a Node.js framework
- 161 (<u>https://nodejs.org/</u>). User authentication is managed using a Keycloak server
- 162 (<u>https://www.keycloak.org/</u>), which supports a single sign-on for users using credentials from connected
- 163 RGD databases.
- 164 A matchmaking query using the OSMP starts with the user defining a gene of interest, the maximum
- allele frequency of the returned variants (with a maximum value of 0.05), the RGD database(s) to be
- 166 queried, and genome reference build in which to display the results. The OSMP sends the specified
- 167 query to the selected database(s) and returns a table of variant occurrences (VOs, defined as a given
- variant in a given participant) meeting the specified criteria. The University of California Santa Cruz
- 169 (UCSC) LiftOver tool (<u>https://genome.ucsc.edu/cgi-bin/hgLiftOver</u>) converts genomic coordinates of
- 170 returned VOs to the user's specified genome reference build, if necessary.

171 Each row of the result table corresponds to a specific VO, however, the rows can also be consolidated to 172 display one unique variant per row. The results table is comprised of three categories of information 173 returned for each VO, with multiple contributing data sources. The first category, variant information 174 (Fig 1a), describes each variant, including its chromosome, genomic coordinates, reference allele, and 175 alternate allele. This information is obtained from the PhenoTips® variant store, an indexed database 176 sourced from variant files for each participant in the PhenoTips<sup>®</sup> instance. The second category, variant 177 annotations (Fig 1b), involves more detailed variant information including predicted changes to the 178 cDNA and amino acid sequences, the allele frequency in the gnomAD control database, and predictions 179 of pathogenicity using the in-silico algorithms from Combined Annotation Dependent Depletion (CADD)<sup>16</sup> and SpliceAl<sup>17</sup>. OSMP annotates this information at the time the query is returned using recent 180 gnomAD<sup>18</sup> and CADD<sup>16</sup> annotations. This "on the fly" annotation was important to harmonize variant-181 182 level annotations within and between databases to accommodate the querying of source GWS files that 183 may have been processed using different bioinformatic pipelines or at different times. This annotation 184 step is currently performed for genes less than 200,000 bps in size due to computing limitations of the 185 beta version of this platform. The final category of information, participant-level details (Fig 1c), includes 186 information specific to the participant in whom each VO is found, including the zygosity of the variant, 187 the clinical features in the form of a standardized vocabulary known as the Human Phenotype Ontology (HPO),<sup>19</sup> and previously identified candidate genes. This information is extracted from the participant's 188 189 PhenoTips<sup>®</sup> record for each queried RGD database. The OSMP calculates some information on the fly 190 including the number of heterozygous and homozygous participants returned for each VO (across all 191 queried database(s)), as well as the number of VOs a participant has across the query gene (termed 192 "burden"). Burden is especially important for identifying participants with potentially compound 193 heterozygous VOs. A more detailed description of each data column returned or calculated by the OSMP 194 is available in Additional File 1.

The results table displayed by the OSMP is flexible, allowing users to customize the interface to best fit their workflow. Columns can be rearranged within the table or hidden, and each column can be filtered independently to narrow down to a list of VOs of interest. These filters are applied through the user's local internet browser, meaning they are instantaneous and do not require an OSMP query to be rerun if filters are changed.

## 200 Participant population

The beta version of the OSMP is connected to a single database, Genomics4RD,<sup>20</sup> that houses data from 201 202 thousands of participants enrolled in the Care4Rare Canada RGD gene discovery research program.<sup>21</sup> 203 Table 1 provides an overview of the participant demographics within Genomics4RD at the time of each 204 of the two OSMP pilot studies. These participants include individuals who are affected with a RGD, many 205 of whom are undiagnosed, as well as their unaffected family members. Affected participants in 206 Genomics4RD are phenotyped using HPO terms, with participants having between 1 and 79 terms listed. 207 Many of these affected participants presented with neurodevelopmental disorders, with the most common HPO terms across the database being global developmental delay (HP:0001263), seizures 208 209 (HP:0001250), delayed speech and language development (HP:0000750), short stature (HP:0004322), 210 and generalized hypotonia (HP:0001290). Numerous participants had one or more genes flagged in their 211 record, classified as either as candidates, i.e., their role in the participant's condition was inconclusive, 212 or causal, i.e., they were a likely explanation for some or all of the participant's presentation. Finally, 213 participants had GWS data uploaded to their records, which had been aligned and annotated using the 214 Care4Rare bioinformatics pipeline.<sup>22</sup>

215

### Table 1 Genomics4RD participants at the time of each pilot study.

	2022 OMIM gene pilot	2023 Novel gene pilot
Total participants, n	2,503	4,063
Unaffected participants, n (%)	1,152 (46%)	1,934 (48%)
Affected participants, n (%)	1,351 (54%)	2,129 (52%)
With 1 or more identified candidate or	618 (46%)	791 (37%)
causal genes, n (%)		

217

## 218 A workflow for one-sided matchmaking

219 One-sided matchmaking on a gene level is expected to result in many potential matches, since it will

return all VOs that exist for a given gene. We therefore devised a workflow to filter the large number of

returned VOs down to a smaller, more manageable list of those most likely to be disease-causing (Fig 2).

222 This seven-step workflow aimed to prioritize VOs that would be most likely to impact protein function

and fit the inheritance pattern of a queried gene in participants with phenotypic overlap with the

224 condition of interest.

225 The first applied filter was based on predicted protein function (i.e., the **consequence filter**). VOs that

were predicted by Ensembl's Variant Effect Predictor (VEP)<sup>23</sup> to have a low impact on protein behaviour

227 (i.e., variants located outside of any exon or splice site) and had a SpliceAI score less than 0.5 were

filtered out. VOs that were not given a conclusive annotation by VEP bypassed this filter and proceeded

to the next step.

230 Next, VOs seen at least once in a homozygous state in gnomAD were removed (i.e., gnomAD hom filter),

as homozygous variants present in reportedly unaffected participants are unlikely to be disease-causing,

regardless of the inheritance pattern for a disease-gene association.

At the third stage of the workflow, we defined three streams based on different inheritance patterns for

a given disease-gene association: an autosomal dominant (AD) stream, an X-linked recessive (XLR)

235 stream, and an autosomal recessive (AR) stream. In each stream, a zygosity filter was first applied, 236 meaning that only heterozygous VOs were kept in the AD stream, hemizygous VOs were kept in the XLR 237 stream and homozygous, or occurrences of multiple heterozygous VOs in the same participant, were 238 kept in the AR stream. For VOs in genes associated with AD disease with a severe pediatric onset, we 239 applied an additional filter to remove variant occurrences present in any zygosity in gnomAD (i.e., 240 gnomAD het filter), since this control database is expected to be largely absent of severe pediatric disease.<sup>18</sup> 241 242 Next, we used the queried genomic database as a control cohort, similar to how the gnomAD filters 243 were used. If a VO was seen in an unaffected participant with the inheritance pattern anticipated for the 244 disease-gene association, the VO was filtered out (i.e., affected status filter). For example, when 245 querying an AR disease-gene association, a homozygous VO in an unaffected participant would result in 246 the VO being filtered out for them and any other participant homozygous for that same variant. 247 Finally, we were left with a prioritized list of VOs with which to perform phenotype/genotype correlation 248 with the original disease-gene query. The HPO terms of each participant with a remaining VO were 249 compared to the clinical descriptions of the disease being queried by a certified genetic counsellor 250 (Author MO). VOs in participants with insufficient phenotype overlap with gueried disease features 251 were removed (i.e., **phenotype filter**), and remaining VOs were then manually reviewed using external 252 information sources (i.e., external data filter). Such information included ClinVar variant classifications,

detailed clinical notes, and sequencing data from family members not included in the RGD database. The
VOs remaining following this final filter were reviewed by a multidisciplinary team including medical
geneticists, laboratory geneticists, and genetic counsellors to determine if they warranted clinical

256 validation or further investigation.

257 Results

# 258 One-Sided Matchmaking Pilot – OMIM Genes

259	Our first pilot of the one-sided matchmaking workflow using the beta version of the OSMP involved
260	querying a set of genes recently associated with disease in OMIM (i.e., OMIM gene pilot). As these
261	disease-gene associations would not have been documented at the time of clinical GWS analysis, we
262	hypothesized this gene set may be enriched for disease-causing variants in unsolved RGD patients in the
263	Genomics4RD database. First, a list of new disease-gene associations added to OMIM between
264	November 2021 and August 2022 was generated (n=227). These associations were then narrowed down
265	to genes with sizes that fit within the OSMP's current memory limitation of 200,000bps. Finally, the
266	remaining associations were manually reviewed to prioritize diseases with presentations most relevant
267	to the Genomics4RD patient population (i.e., diseases that severely impact a single system, or multiple
268	systems), resulting in a final list of 130 disease-gene associations (across 116 unique genes, i.e., 14 genes
269	had two newly described disease associations) to be queried using the OSMP. Approximately 63%
270	(82/130) of the disease-gene associations were for AR conditions, 32% (42/130) for AD conditions, and
271	5% (6/130) for XLR conditions.
272	In total, 12,872 VOs with an allele frequency of 0.01 or less in Genomics4RD were returned by the OSMP
273	across the 130 disease-gene associations queried. Figure 2 and Table 1 detail the number and
274	proportions of VOs removed at each stage of the workflow, respectively. The consequence filter
275	removed 35% of the VOs with a VEP-annotated impact. Next, filtering out VOs seen in a homozygous
276	state in gnomAD removed 27% of remaining VOs. The effectiveness of the zygosity filter varied across
277	the inheritance patterns, with 11% of the remaining VOs filtered out VOs for AD associations, 51% of
278	VOs for XLR associations, and 94% of VOs for AR disease-gene associations. For early-onset AD
279	conditions, 68% of the VOs that passed the zygosity filter were removed due to their presence in the
280	gnomAD database. An additional 79% of the VOs for AD disease-gene associations, 83% for XLR
281	associations, and 74% of VOs for AR associations were removed as they were present in the anticipated

282	zygosity in unaffected participants. This resulted in 305 VOs remaining after all filters using OSMP-
283	provided data. Across all inheritance patterns, 81% of the remaining VOs were filtered out due to
284	insufficient phenotypic overlap with the disease synopsis in OMIM. Finally, 46% of the remaining VOs
285	were removed following a manual review using data sources external to the OSMP. We filtered out VOs
286	at this stage due to ClinVar classifications as benign or likely benign, variants not segregating
287	appropriately in family members, and insufficient phenotypic overlap with the OMIM disease following
288	review of external clinical notes. Following all filtration steps, a total of 31 VOs (0.24% of the original
289	query results) remained across 20 newly described disease-gene associations and were prioritized for
290	review with the multidisciplinary team: 70% (14/20) of these disease-gene associations were for
291	autosomal dominant conditions, and 30% (6/20) were for autosomal recessive conditions. Of these, one,
292	so far, has resulted in a diagnosis. A de novo VO in the gene POLR3B was identified in a patient with
293	overlapping neurological features to the recently described association with demyelinating Charcot-
294	Marie-Tooth disease type 1I (OMIM 619742). This variant, identified through the OMIM gene pilot, has
295	since been classified as likely pathogenic through validation by a clinical diagnostic laboratory.

# **Table 2. Efficacy of workflow filters in the OMIM gene and novel gene pilots.** Percentages indicate the proportion of remaining VOs removed

by each filter step.

Filter Name	Filter Description	OMIM gene pilot			Novel gene pilot				
Consequence filter	Removes low impact VOs, unless predicted to impact splicing	35%			28%				
gnomAD hom filter	Removes VOs homozygous in >0 gnomAD samples		27	'%		25%			
7	Removes VOs where zygosity	AD	AD XLR AR		AD		XLR	AR	
Zygosity filter	inheritance	11%		51%	94%	3%		65%	91%
gnomAD het filter	Removes VOs heterozygous in	Severe Pediatric	Other	N/A	N/A	Severe Pediatric	Other	N/A	N/A
		68%	N/A			75%	N/A		
Affected status filter	Removes VOs with expected inheritance in >0 unaffected participants	79%		83%	34%	82%		45%	87%
Phenotype filter	Removes VOs with insufficient phenotype overlap with queried disease features	81%			91%				
External data filter	Removes VOs based on information external to OSMP	46%			67%				

### 299 One-Sided Matchmaking Pilot – Novel Genes

300 The second one-sided matchmaking pilot queried a set of candidate genes assigned to a disease-gene 301 association in OMIM (designated as the novel gene pilot). These candidate genes were identified 302 through the analysis of GWS data for unsolved RGD patients enrolled in the Care4Rare research program (criteria described by Osmond et al).<sup>8</sup> Most of these participants have records in the Genomics4RD 303 304 database and are gueriable via the OSMP. This second pilot had two goals: first, to validate the one-305 sided matchmaking workflow (i.e., do we identify the true positive participants who harbor compelling 306 candidates in these genes), and second to identify additional families with rare variants in the same 307 gene and overlapping phenotype that would help to build evidence for a novel disease-gene association. 308 A list of 178 novel candidate genes was queried using the OSMP for this novel gene pilot. Approximately 309 62% (111/178) of these genes were hypothesized to be associated with an AD condition, 33% (58/178) 310 were hypothesized to be associated with an AR condition, and 5% (9/178) were hypothesized to be 311 associated with an XLR condition. Of these 178 candidate genes, 140 had true positive participants with 312 data accessible to the OSMP.

#### 313 Validation of the one-sided matchmaking workflow

314 We queried OSMP with the set of 140 genes with VOs previously prioritized as disease-causing or strong 315 candidates in Care4Rare families to test the one-side matchmaking workflow. After applying all filters, 316 the VOs from 89% (124/140) of these genes remained within our prioritized list. For the 16 genes with 317 VOs that did not pass all filters, six genes had VOs that were removed by the consequence filter, most 318 commonly because the VOs occurred just outside a canonical splice site. Nine had VOs removed by the 319 gnomAD het filter, as these genes were associated with a severe AD pediatric onset condition, but the 320 VO was present in at least one individual in gnomAD. In all these cases, the presence of the variant in 321 gnomAD had been previously identified and the variant was considered a weak novel candidate. Finally,

322 one gene had a VO that was removed by the **affected status filter**, as the variant was present in a family 323 member marked as 'unaffected', however on reflection it was deemed that the affected status of this 324 relative was inconclusive. Upon review, we did not believe that these false negatives warranted changes 325 to our one-sided matchmaking protocol.

326 Use of OSMP to identify additional novel candidate gene families

327 We excluded the true positive families described above in our candidate gene query of OSMP for 328 additional families with seemingly the same novel RGD. In total, 20,308 VOs with a maximum allele 329 frequency of 0.01 in Genomics4RD were returned by the OSMP related to the 178 candidate genes. 330 Figure 3 and Table 2 detail the number and proportions of VOs removed at each stage of the workflow, 331 respectively. Overall, the consequence filter removed 28% of eligible VOs, and the gnomAD het filter 332 removed 25% of the remaining VOs. Like the OMIM gene pilot, the efficacy of the zygosity filter differed 333 between the hypothesized inheritance patterns for these novel genes. Approximately 3% of VOs for AD 334 genes, 65% of VOs for XLR genes and 91% of VOs for AR genes were removed by this filter. Filtering out 335 variants seen in gnomAD removed 75% of the remaining VOs from genes with a suspected early-onset 336 AD condition. When filtering out VOs seen in unaffected participants, 82% of the VOs for AD disease-337 gene associations, 45% for XLR associations and 87% for AR associations were removed. This resulted in 338 604 VOs remaining after all filters using OSMP-provided data. Filtering out participants with insufficient 339 phenotype overlap to the original patient in which the novel gene was identified resulted in the removal 340 of 91% of VOs across all patterns of inheritance. Finally, approximately 67% of the remaining VOs were 341 removed using data not available directly through the OSMP. External data used to rule out these VOs 342 was similar to the OMIM gene pilot and included sequencing data from family members not in 343 Genomics4RD, more extensive notes on clinical presentation, and the number of hemizygotes who carry 344 a variant in gnomAD for XLR genes. Following all filtration steps, a total of 18 VOs (0.09% of the initial 345 query results) across 14 novel candidate genes remained for review with the multidisciplinary team. Ten

346 of these novel candidate genes were hypothesized to be associated with AD conditions (10 347 heterozygous VOs), three genes were hypothesized to be associated with AR conditions (2 homozygous 348 VOs, and 2 heterozygous VOs in the same participant), and one gene was thought to be associated with 349 a XL condition (1 hemizygous VO). One of these remaining VOs is in the gene CCDC6 and it is believed to 350 be the molecular cause for this participant's rare disease. This heterozygous CCDC6 variant has already 351 been described in several other individuals in an ongoing collaboration first established through the 352 MME, and this participant represents an addition to this cohort. Review of the other prioritized VOs is 353 ongoing.

## 354 Discussion

The development and piloting of the beta version of the OSMP shows that one-sided matchmaking can be effective in identifying genetic variants of interest in undiagnosed patients with RGD. Though tens of thousands of VOs were returned in each pilot, we developed an effective workflow to quickly filter VOs to those most likely to be disease-causing. While clinical review of the prioritized VOs is ongoing, there has been one diagnosis made in the newly described OMIM gene *POLR3B*, and one diagnosis made in the novel gene *CCDC6* where collaboration is ongoing. We anticipate that further review of prioritized VOs will lead to diagnoses in additional genes.

The OMIM gene and novel gene pilots highlight the importance of making participant-level information available when performing this type of matchmaking to rule out as many false positives as possible before undergoing more extensive case reviews. Knowing when variants are present in unaffected participants was highly effective in filtering VOs across genes associated with all inheritance patterns, in total removing 78% and 79% of the remaining VOs in the OMIM gene and novel gene pilots, respectively. Similarly, access to phenotypes in the form of HPO terms for affected participants enabled the removal of over 80% of the remaining VOs across both one-sided matchmaking pilots. Lastly,

369 zygosity information on VOs was especially effective as a filter for genes with a known or hypothesized 370 AR inheritance pattern, resulting in the removal of over 90% of remaining VOs for both pilot studies. The 371 utility of phenotypic and genotypic data in proactively ruling out potential matches is a trend that our 372 team has also experienced with two-sided matchmaking - over half of two-sided matches were ruled out 373 when such information was available at the time the initial match was made.<sup>8</sup> Increasing the inclusion of 374 such participant-level information for matches submitted to the Matchmaker Exchange has been highlighted as an important factor to improve the efficiency of two-sided matchmaking,<sup>7</sup> and our pilots 375 376 suggest that one-sided matchmaking platforms will not be successful without this data being made 377 available. 378 While the beta version of the OSMP can provide the information necessary to make efficient one-sided 379 matchmaking queries through an interface that is easy to use and customizable, these pilot studies 380 provide insight into ways that the platform can be further improved for more widespread use. Providing

381 on the fly annotations with CADD and gnomAD datasets ensures variant-level data returned by the

382 OSMP remains accurate and harmonized across databases using different bioinformatics pipelines,

however this feature is currently limited to genes less than 200,000 base pairs in size. Allocation of

383

392

optimized for family-based analyses.

384 additional computing resources or performance improvements to the existing software may be required 385 to ensure this feature is available to queries of genes of all sizes. Reviewing the VOs that were ruled out 386 using data not directly available through the OSMP also indicates ways in which the platform can be 387 improved. Knowledge of whether a variant has been classified in ClinVar, or if variants in XLR genes are 388 seen in a hemizygous state in gnomAD, would both be useful to have available in the OSMP annotations. 389 More detailed information on the inheritance of VOs (i.e., if a heterozygous variant is de novo vs 390 inherited, or if multiple heterozygous variants are in cis or trans) would also improve the efficiency of 391 one-sided matchmaking on this platform, as the results generated by the OSMP are not currently

393 Finally, future versions of the OSMP will be expanded in both the number of users with access to the 394 platform, and the number of connected databases. Upcoming OSMP development will focus on load 395 testing the tool so that more users – both members of the Genomics4RD database and other third-party 396 researchers - can utilize this resource. In this next phase, connecting additional databases to the OSMP 397 will be crucial for both increasing the number of unsolved RGD patients made available for one-sided 398 matchmaking and improving the OSMP's ability to rule out existing VOs of interest with an increased 399 number of internal controls (i.e., unaffected family members). The OSMP is in the process of establishing 400 a connection to the PhenoTips<sup>®</sup> database maintained by the Genomic Answers for Kids rare disease 401 research program, which will enable the OSMP to query over 2,900 additional RGD participants with 402 GWS data.

#### 403 Conclusions

404 The beta version of the OSMP can perform gene-level one-sided matchmaking queries for the purposes 405 of prioritizing variants of interest in undiagnosed RGD patients. The development and piloting of the 406 one-sided matchmaking workflow for both newly described disease-gene associations and novel 407 candidate genes demonstrates both the sheer number of variant occurrences returned by gene-level 408 queries, and the importance of variant-level and participant-level data in filtering possible matches 409 down to a level more reasonable for users to review. Further, our pilots act as proof-of-principle that 410 one-sided matchmaking can identify additional diagnoses and candidate genes. The lessons learned 411 from piloting the beta version of the OSMP will be used to further refine the functionality of the 412 platform, and we believe these insights will be of use to other groups developing similar tools. The 413 connection of additional RGD databases to one-sided matchmaking services like the OSMP will be crucial in providing access to matchmaking for as many clinically undiagnosed RGD families as possible, in the 414 415 hopes of identifying the genetic etiologies of their conditions.

## 416 List of abbreviations

- 417 **AD:** Autosomal dominant
- 418 **AR:** Autosomal recessive
- 419 CADD: Combined Annotation Dependent Depletion
- 420 CMA: Chromosomal microarray
- 421 **DECIPHER:** DatabasE of genomiC variation and Phenotype in Humans using Ensembl Resources
- 422 FORGE: Finding of Rare Disease Genes
- 423 **GPAP:** Genome-Phenome Analysis Platform
- 424 GWS: Genome wide sequencing
- 425 *HPO:* Human Phenotype Ontology
- 426 MME: Matchmaker Exchange
- 427 OMIM: Online Mendelian Inheritance in Man
- 428 **OSMP:** One-Sided Matching Portal
- 429 *RGD:* Rare genetic disease
- 430 UCSC: University of California Santa Cruz
- 431 VEP: Variant Effect Predictor
- 432 VO: Variant occurrence
- 433 XLR: X-linked recessive
- 434

### 435 Declarations

#### 436 Ethics approval and consent to participate

- 437 Informed consent was obtained for all participants that were queried by the OSMP in this study.
- 438 Participant-level data made available for the purposes of matchmaking was de-identified. Institutional
- 439 review board approval was obtained from the Children's Hospital of Eastern Ontario for both Finding of
- 440 Rare Disease Genes in Canada (FORGE) and Care4Rare Canada (Research Ethics Board # 11/04E).
- 441 Institutional review board approval for Care4Rare-SOLVE was obtained from Clinical Trials Ontario
- 442 (CTO1577).

#### 443 Consent for publication

444 Not applicable.

#### 445 Availability of data and materials

- 446 The source code for the OSMP is available at <u>https://github.com/ccmbioinfo/osmp</u>. Access to the beta
- 447 version of the OSMP is currently limited to a small set of Genomics4RD users. The genotypic and
- 448 phenotypic data that support the findings of this study are located in the controlled access database
- 449 Genomics4RD. Genomics4RD open access data is available at <a href="https://www.genomics4rd.ca/">https://www.genomics4rd.ca/</a>.

## 450 Competing interests

- 451 OJB and MB have an equity interest in, and OJB is an employee of PhenoTips<sup>®</sup>, which licenses software
- 452 used in the Genomics4RD database.
- 453 Funding

454 This study was funded through the *Genomics and Precision Health Top-up* grant GPT-174518

- 455 "Care4Rare-Solve: Efficient cross-border matchmaking to deliver diagnoses for rare genetic diseases",
- awarded by the Canadian Institutes of Health Research.
- 457 The development of Genomics4RD and the production of its housed GWS data was performed under the
- 458 Care4Rare Canada Consortium funded by Genome Canada and the Ontario Genomics Institute (OGI-
- 459 147), the Canadian Institutes of Health Research, Ontario Research Fund, Genome Alberta, Genome
- 460 British Columbia, Genome Quebec, and Children's Hospital of Eastern Ontario Foundation.
- 461 T.H. was supported by a Frederick Banting and Charles Best Canada Graduate Scholarship Doctoral
- 462 Award from Canadian Institutes of Health Research. MB is a CIFAR AI Chair. KMB was supported by a
- 463 Canadian Institutes of Health Research Foundation grant FDN-154279 and a Tier 1 Canada Research
- 464 Chair in Rare Disease Precision Health. The work at Children's Mercy Kansas City is supported by
- 465 generous donors to Children's Mercy Research Institute and Genomic Answer For Kids program.

#### 466 Authors' contributions

- 467 Funding acquisition for this study was conducted by EMP, TH, TP, MB, and KMB. MO, EMP, OJB, TH, CK,
- 468 ALT, TP, MB, and KMB contributed to the conceptualization and high-level design of OSMP.
- 469 Development and testing of OSMP was performed by MO, EMP, OJB, MF, MC, CK, HGBHL, JX, DS, KL,
- 470 KM, HW, HGD, ALT, and MB. Curation of the data to be queried by OSMP was conducted by EMP, MC,
- 471 AJ, HW, HGD, WC, AC, EF, and IT. MO, EMP, and TH designed the one-sided matchmaking workflow.
- 472 MO, HW, and EW collected the data generated from the OSMP pilot studies. Data analysis from the pilot
- 473 studies was performed by MO and EMP. The initial manuscript draft was written by MO, EMP, and KMB.
- 474 All authors have read and approved the final manuscript.

# 475 Acknowledgements

- 476 We would like to acknowledge Alina Gvozdik, John Q. Miller, Joel John, Kevin M. Power, John N. Gregor,
- 477 and Bourke B. Hutchison for their contributions towards establishing API connections between the
- 478 OSMP and its connected databases.

# 479 References

480	1.	Clark MM, Stark Z, Farnaes L, et al. Meta-analysis of the diagnostic and clinical utility of genome
481		and exome sequencing and chromosomal microarray in children with suspected genetic
482		diseases. NPJ Genom Med. 2018;3(1):1-16. <u>https://.doi.org/10.1038/s41525-018-0053-8</u>
483	2.	Miller DT, Adam MP, Aradhya S, et al. Consensus Statement: Chromosomal Microarray Is a First-
484		Tier Clinical Diagnostic Test for Individuals with Developmental Disability or Congenital
485		Anomalies. Am J Hum Genet. 2010;86(5):749-764. <u>https://.doi.org/10.1016/j.ajhg.2010.04.006</u>
486	3.	Tan NB, Stapleton R, Stark Z, et al. Evaluating systemic reanalysis of clinical genomic data in rare
487		disease from single center experience and literature review. Mol Genet Genomic Med.
488		2020;8(11):1-19. <u>https://doi.org/10.1002/mgg3.1508</u>
489	4.	Hartley T, Soubry É, Acker M, et al. Bridging clinical care and research in Ontario, Canada:
490		Maximizing diagnoses from reanalysis of clinical exome sequencing data. Clin Genet.
491		2023;103(3):288-300. <u>https://doi.org/10.1111/cge.14262</u>
492	5.	Bamshad MJ, Nickerson DA, Chong JX. Mendelian Gene Discovery: Fast and Furious with No End
493		in Sight. Am J Hum Genet. 2019;105(3):448-455. <u>https://doi.org/10.1016/j.ajhg.2019.07.011</u>
494	6.	Strande NT, Rooney Riggs E, Buchanan AH, et al. Evaluating the Clinical Validity of Gene-Disease
495		Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. Am J
496		Hum Genet. 2017;100(6):895-906. <u>https://doi.org/10.1016/j.ajhg.2017.04.015</u>
497	7.	Boycott KM, Azzariti DR, Hamosh A, Rehm HL. Seven years since the launch of the Matchmaker
498		Exchange: The evolution of genomic matchmaking. <i>Hum Mutat</i> . 2022;43(6):659-667.
499		https://doi.org/10.1002/humu.24373
500	8.	Osmond M, Hartley T, Dyment DA, et al. Outcome of over 1500 matches through the
501		Matchmaker Exchange for rare disease gene discovery: The 2-year experience of Care4Rare

502 Canada. *Genet Med*. 2022;24(1):100-108. <u>https://doi.org/10.1016/j.gim.2021.08.014</u>

- 503 9. Stark Z, Dolman L, Manolio TA, et al. Integrating Genomics into Healthcare: A Global
- 504 Responsibility. *Am J Hum Genet*. 2019;104(1):13-20. <u>https://doi.org/10.1016/j.ajhg.2018.11.014</u>
- 505 10. Cranage A. Our UK Biobank Journey: 3 years and over 240,000 human genomes. Wellcome
- 506 Sanger Institute. September 16, 2022. Accessed August 15, 2023.
- 507 https://sangerinstitute.blog/2022/09/26/our-uk-biobank-journey-3-years-and-over-240000-
- 508 <u>human-genomes/</u>
- 509 11. Rodrigues ES, Griffith S, Martin R, et al. Variant-level matching for diagnosis and discovery:
- 510 Challenges and opportunities. *Hum Mutat*. 2022;43(6):782-790.
- 511 <u>https://doi.org/10.1002/humu.24359</u>
- 512 12. Foreman J, Brent S, Perrett D, et al. DECIPHER: Supporting the interpretation and sharing of rare
- 513 disease phenotype-linked variant data to advance diagnosis and research. *Hum Mutat*.
- 514 2022;43(6):682-697. <u>https://doi.org/10.1002/humu.24340</u>
- 515 13. Laurie S, Piscia D, Matalonga L, et al. The RD-Connect Genome-Phenome Analysis Platform:
- 516 Accelerating diagnosis, research, and gene discovery for rare diseases. *Hum Mutat*.
- 517 2022;43(6):717-733. <u>https://doi.org/10.1002/humu.24353</u>
- 518 14. Pais LS, Snow H, Weisburd B, et al. *seqr*: A web-based analysis and collaboration tool for rare
- 519 disease genomics. *Hum Mutat*. 2022;43(6):698-707. <u>https://doi.org/10.1002/humu.24366</u>
- 520 15. Girdea M, Dumitriu S, Fiume M, et al. PhenoTips: patient phenotyping software for clinical and
- 521 research use. *Hum Mutat*. 2013;34(8):1057-1065. <u>https://doi.org/10.1002/humu.22347</u>
- 522 16. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness
- of variants throughout the human genome. *Nucleic Acids Res*. 2019;47(D1):D886-D894.
- 524 <u>https://doi.org/10.1093/nar/gky1016</u>
- 525 17. Jaganathan K, Panagiotopoulou SK, McRae JF, et al. Predicting Splicing from Primary Sequence
- 526 with Deep Learning. *Cell*. 2019;176(3):535-548. <u>https://doi.org/10.1016/j.cell.2018.12.015</u>

- 527 18. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from
- 528 variation in 141,456 humans. *Nature*. 2020;581:434-443. <u>https://doi.org/10.1038/s41586-020-</u>
- 529 <u>2308-7</u>
- 530 19. Köhler S, Gargano M, Matentzoglu N, et al. The Human Phenotype Ontology in 2021. *Nucleic*
- 531 Acids Res. 2021;49(D1):D1207-D1217. https://doi.org/10.1093/nar/gkaa1043
- 532 20. Driver H, Hartley T, Price EM, et al. Genomics4RD: An integrated platform to share Canadian
- 533 deep-phenotype and multi- omic data for international rare disease gene discovery. *Hum Mutat*.
- 534 2022;43(6):800-811. <u>https://doi.org/10.1002/humu.24354</u>
- 535 21. Boycott KM, Hartley T, Kernohan KD, et al. Care4Rare Canada: Outcomes from a decade of
- network science for rare disease gene discovery. *Am J Hum Genet.* 2022;109(11):1947-1959.
- 537 <u>https://doi.org/10.1016/j.ajhg.2022.10.002</u>
- 538 22. Centre for Computational Medicine. Crg2: Clinical research pipeline. Github. Updated July 25,
- 539 2023. Accessed August 15, 2023. https://github.com/ccmbioinfo/crg2
- 540 23. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*.
- 541 2016;17(1):122-136. <u>https://doi.org/10.1186/s13059-016-0974-4</u>

## 542 Figure legends

- 543 Fig. 1. The One-Sided Matchmaking Platform (OSMP). Queries made for a gene of interest return data
- 544 from three sources: (a) Variant information basic information on variants identified in the PhenoTips®
- variant store. (b) Variant annotations extracted from CADD and gnomAD datasets. (c) Participant-level
- 546 details phenotypic and genotypic details from individual PhenoTips<sup>®</sup> record

- 548 Fig. 2. One-sided matchmaking workflow and details of variant occurrences (VOs) returned by the OMIM
- 549 gene pilot queries (n=130 disease-gene associations). Filtration steps are indicated by dark grey boxes,
- and VOs removed by each filter is indicated by light grey boxes. Percentages indicate the proportion of

- 551 VOs from the previous step that were filtered out. AD: autosomal dominant, XLR: X-linked recessive, AR:
- 552 autosomal recessive.
- 553 Fig 3. One-sided matchmaking workflow for variant occurrences (VOs) returned by the novel gene pilot
- queries (n = 178 novel candidate genes). Filtration steps are indicated by dark grey boxes, and VOs
- removed by filters are indicated by light grey boxes. Percentages indicate the proportion of VOs from
- the previous step that were filtered out. AD: autosomal dominant, XLR: X-linked recessive, AR:
- 557 autosomal recessive.
- 558 Additional Files
- 559 Additional file 1.pdf Overview of data columns returned or generated by the OSMP



	*		_		
	Contributors		* Required Field		
	g4rd		Search	Clear	
85 individuals		Customize column	s 🐼 🛛	Export Data	

End ≑	T Ref ≑	T Alt 🗧	Original Assembly ≑	Current Assembly	Source
56524	А	С	GRCh37	GRCh37	g4rd

T cdna ≑	aaChange	▼ consequence ≑	gnomAD_AF 🕕	gnomadHorr	CADD score	SpliceAl score	SpliceAl type
c.1389A>C	p.Leu409Arg	NON_SYNONYMOUS	0	0	29.3	0	NA

ıl	T Family ID ≑	Sex 🖲	Affected Status	T Flagged Gene(s) ≑	Present Phenotypes	Absent Phenotypes	T Ethnicity
					<u>Spasticity</u>		
					<u>Global developmental delay</u>		
	FAM000	Female	Female affected	ABHD16A - solved	Hypoplasia of the corpus callosum		
					Neurodegeneration		
					<u>Joint contracture of the</u> <u>hand</u>		



