Published in final edited form as:

Hum Mutat. 2018 December ; 39(12): 1827–1834. doi:10.1002/humu.23655.

***matchbox*: An open-source tool for patient matching via the Matchmaker Exchange**

Harindra Arachchi^{1,2}, Monica H Wojcik^{1,3}, Benjamin Weisburd¹, Julius O. B. Jacobsen⁴, Elise Valkanas¹, Samantha Baxter¹, Alicia B. Byrne^{1,5,6}, Anne H. O'Donnell-Luria^{1,3}, Melissa Haendel^{7,8}, Damian Smedley⁴, Daniel G. MacArthur^{1,2,9}, Anthony A. Philippakis^{2,*}, and Heidi L. Rehm^{1,2,9,*}

¹Center for Mendelian Genomics, The Broad Institute of MIT and Harvard, Cambridge, MA, USA

²The Broad Institute of MIT and Harvard, Cambridge, MA, USA

³Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

⁴William Harvey Research Institute, Barts & The London School of Medicine & Dentistry, Queen Mary University of London, Charterhouse Square, London, EC1M 6BQ

⁵Department of Genetics and Molecular Pathology, Centre for Cancer Biology, SA Pathology, Adelaide, Australia

⁶School of Pharmacy and Medical Sciences, University of South Australia, Adelaide, Australia

⁷Oregon Clinical and Translational Research Institute, Oregon Health & Science University, Portland, OR, USA

⁸Linus Pauling Institute, Oregon State University, Corvallis, OR, USA

⁹Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

Abstract

Rare disease investigators constantly face challenges in identifying additional cases to build evidence for gene-disease causality. The Matchmaker Exchange (MME) addresses this limitation by providing a mechanism for matching patients across genomic centers via a federated network. The MME has revolutionized searching for additional cases by making it possible to query across institutional boundaries, so that what was once a laborious and manual process of contacting researchers is now automated and computable. However, while the MME network is beginning to scale, the growth of additional nodes is limited by the lack of easy-to-use solutions that can be implemented by any rare disease database owner, even one without significant software engineering resources. Here we describe *matchbox*, which is an open-source, platform-independent, portable bridge between any given rare disease genomic center and the MME network, which has already led to novel gene discoveries. We also describe how *matchbox* greatly reduces the barrier to participation by overcoming challenges for new databases to join the MME.

Corresponding author: Heidi L. Rehm, The Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA, 02142, hrehm@mgh.harvard.edu.

*these authors contributed equally; order determined by roshambo

Keywords

rare disease; Matchmaker Exchange; open-source; novel gene; matchbox

Introduction

The Matchmaker Exchange (MME) (Philippakis et al., 2015), launched in 2013 and reliant on standards defined within the Global Alliance for Genomics and Health, allows investigators searching for novel genetic causes of rare disease to find individual patients with the same candidate gene and overlapping phenotypic features. MME, made up of nine nodes located in five countries and supporting users from over 73 countries, mitigates the challenge of data ownership and international regulatory limitations regarding movement of data across country boundaries, and provides more site control of sharing preferences and matching algorithms by utilizing an application programming interface (API) that allows local hosting of data. Patient data shared through the MME are anonymized and no protected health information (PHI) is shared. Once a match happens between two participant genomic centers, the two “data-owners” of the patients are put into contact with each other to evaluate the validity of the match. There are two routes by which new candidate genes can be submitted into the network: addition by existing nodes or launch of a new node onto the federated network. *matchbox* (<https://github.com/macarthur-lab/matchbox>), a software solution designed to be easily installed in any number of centers worldwide, facilitates the latter route.

When MME was launched, a center (node) interested in joining the MME had to build the software functionality - either from scratch or off of an illustrative open source reference implementation (<https://github.com/MatchmakerExchange/reference-server>) not intended for a production setting. While there were a few other open source implementations (<https://github.com/ga4gh/mme-apis/wiki/Implementations>), these were coupled to their host institution infrastructure and purpose. To enable faster growth of the MME network, we launched *matchbox* to reduce technology as a barrier to launching a new node and thereby provide matchmaking capabilities to many more patients of diverse origins worldwide. While *matchbox* and *seqr* have been previously introduced as part of a protocol for MME use (Sobreira et al., 2017), we now provide more detail on the design, implementation, and usage of these applications, highlighting their abilities to improve the process of novel gene discovery.

Rationale for building *matchbox*

Understanding and building a matchmaker server can be challenging, consuming both time and resources. *matchbox* is a ready-built, scalable, easy-to-use, minimal but robust solution for aiding the launch of a new rare disease center as part of the MME federated network. Although *matchbox* reduces the barriers to entry through the provision of consistent basic infrastructure for matching, the host is still able to improve or change matching algorithms as they deem fit. *matchbox* can handle communication between matchmaker centers and act as a repository that is isolated from a center’s general patient database, containing only cases

whose attributes are allowed to be shared within the network. Currently, it takes about 30 minutes to download, setup, and start an instance on a Unix-like system. *matchbox* is intended for an institution supporting many end-users and with software development capabilities and informatics resources sufficient to host and start a server in a Unix environment. Such institutions tend to have an existing database for storing local patient data from its local users and are looking for a means of connecting that application to the MME, but have limited engineering resources to build such a bridge, or would prefer to use an existing solution. However, for those also lacking an existing database, we propose the *seqr* application (<https://github.com/macarthur-lab/seqr>), a companion freely available open source analysis platform that can be easily integrated with *matchbox* (described later in the manuscript).

As of now, *matchbox* is not designed for a single individual user, though it is quite possible for a single user familiar with Unix environments to start and maintain a *matchbox* instance.

matchbox implements the Matchmaker Exchange API specification (Buske et al., 2015), whereby a new node joining the MME has the option of using *matchbox* versus the resource-intensive route of building their own software. Given *matchbox*'s freely available open-source nature, we encourage the community to join in to maintain, augment, and build on this application so that it will lead to continuous improvement and further collaborative open science approaches to rare disease diagnosis.

Matchbox architecture

matchbox is built using Java 1.8 (Oracle Corporation, Redwood Shores, CA) and the Spring boot (<https://projects.spring.io/spring-boot/>) framework (Pivotal Software, San Francisco, CA). It currently uses MongoDB (<https://www.mongodb.com/>) as the primary database. *matchbox* has the following endpoints (see Table). Patient data in *matchbox* was collected under a Partners Healthcare IRB-approved protocol.

Matchbox matching algorithm

The heart of *matchbox* is its matching algorithm which it uses to determine whether two distinct patients are similar to each other. Similarity is quantified to guide prioritization of matches. As per MME API specification, each result pair must provide an enumeration between 0 and 1 that represents the similarity strength and each host institution determines the algorithm used for matching. *matchbox* separately scores on genotype and phenotype similarity, followed by the merging of these into a single overall similarity score between 0 and 1 that we return with the result as specified by the MME API. We also return the individual genotype and phenotype scores that contributed to the merged score for transparency and data richness.

The extent of similarity of each result pair is quantified by:

$$\text{when genotype score} \neq 0 \text{ and phenotype score} \neq 0 \Rightarrow \text{result score} = \text{genotype score} \times \text{phenotype score}$$

It must be highlighted that the MME operates in an imperfect universe with varying amounts of data shared per patient query. A great part of the success of MME has been its capability to be flexible and allow queries to contribute as much data as possible above a bare minimum, such as simply gene names and or phenotypes. While it is understood that the strength of the result is largely contingent upon how descriptive the query is, the algorithm must be able to function at both ends of the data-sharing spectrum. To allow for this spectrum of information availability, we have added in pivots that prevent the system from refusing queries and adjust itself as appropriate to data provided.

The genotype scoring strategy

When genotype data are provided, they are used as the primary matching criteria, with phenotype data used in scoring to further rank the genotype matches. *matchbox* does have the capability to match *only* on phenotypes in the absence of genotype or when the given genotypes result in no matches. The host center can decide whether or not to utilize this latter feature as phenotype-only matching may not be universally desired. We have observed that default inclusion of matching *only* on phenotypes is often undesirable by users given the level of false positives that it generates; we believe this will improve over time with better phenotyping.

Our base criteria for a match is two patients containing the same candidate gene, with or without defined variant(s). The similarity of genotypes is calculated using constraint scores, population frequencies via gnomAD (<http://gnomad.broadinstitute.org>) and the patient population within *matchbox*. To optimize performance of this calculation and minimize network travel, we cache gnomAD metrics.

As the first step of our matching process, the candidate gene +/- specific variant(s) for each patient in *matchbox* are compared pairwise with the query patient. When genes in common are noted, a baseline metric is generated with the help of gnomAD exome data. If a specific variant is provided it is used to fetch the exome allele frequency from the gnomAD API (<http://gnomad-api.broadinstitute.org>). For example, we use the start and stop positions, chromosome, alternate and reference bases using a query similar to,

```
{ {variant(id:"7-64438667-G-A",source:"exome") }
```

We then calculate the inverse of this allele frequency. The idea behind this inversion is that if a variant is present at an appreciable allele frequency in a normal population (a population depleted of severe, pediatric-onset disease), it is less likely to be disease-causing.

$$\text{If gnomad exome allele frequency} \neq 0 \Rightarrow \text{gnomad score of gene candidate} = (\text{gnomad exome allele frequency})^{-1}$$

When incomplete variant data are provided or the allele frequency is absent in gnomAD; a constraint-based (Lek et al., 2016) score is used instead. This score is determined using the loss-of-function (LoF), missense, and synonymous constraint Z-scores as well as pLI values

that are retrieved from the gnomAD API based on the gene ID (preferably Ensembl ID) using a query similar to,

$$\{gene(gene_id: "ENSG00000152926") \{exacv1_constraint \{pLI\} \{syn_z\} \{mis_z\} \}\}$$

If the variant type is specified in the patient record, then the score is set to the missense Z-score for missense variants, and synonymous Z-score for synonymous variant. For predicted loss of function or an unsupported variant type, pLI is used. When no type is given, as it is an optional value, the missense constraint Z score is used.

If gnomad exome allele frequency = 0 || incomplete variant information $\Rightarrow$ gnomad score of gene candidate = (gene constraint based Z score)

If a patient has more than one candidate gene specified, we use their score product. Currently there are no means of weighing candidate genes based on importance; hence we assume each has equal importance in disease; we do plan to introduce weights in the future as discussed later.

$$gnomad\ score = score\ of\ gene_candidate_1 \times score\ of\ gene_candidate_2$$

The resulting gnomAD based score is then scaled to between 1 and 0 using the logistic function and is used as the baseline genotype score.

Once all patients in *matchbox* that satisfy this minimum criteria of having a candidate gene in common are identified and a baseline score is calculated, we further augment this score by calculating the prevalence of matches in *matchbox* (e.g. have a variant in the same gene as the query with either the same type, zygosity or variant position), and multiply the baseline gnomAD score with this percentage. If there is more than one match the probability of it being disease-causing is higher.

$$matchbox\ score = (number\ of\ matches\ in\ matchbox \div total\ number\ of\ patients\ in\ matchbox)$$

Thereby the final genotype score arises from:

$$base\ genotype\ score = gnomad\ score \times matchbox\ score$$

The phenotype matching strategy

matchbox uses the Exomiser (Smedley et al., 2015) Phenodigm (Smedley et al., 2013) phenotype matching module available from the Monarch Initiative (Mungall et al., 2017) to calculate the similarity between sets of phenotypes between two patients. The Phenodigm algorithm compares all query phenotypes encoded using the Human Phenotype Ontology (HPO) (Kohler et al., 2017) against the set of all HPO disease profiles to produce the best possible match. Then all node phenotype profiles are compared pairwise against the query phenotypes and the resulting best matches are compared against the perfect match to

produce a final percentage score. The pairwise scores are calculated as the geometric mean of the jaccard similarity of the two pairs and the information content of their lowest common subsumer (Kohler et al., 2017). Additional details of this algorithm have been previously published (Bone et al., 2016; Robinson et al., 2014; Smedley et al., 2014; Smedley et al., 2015; Smedley et al., 2016; Zemojtel et al., 2014).

matchbox* integration with *seqr

matchbox was built as a backend server that facilitates communication with the MME and does not have a native graphical user interface. We use the *seqr* web application for this purpose (<https://seqr.broadinstitute.org/>). *seqr* is the analytical platform for the Center for Mendelian Genomics (CMG) at the Broad Institute and is open source software. Our collaborators around the world use the instance hosted at Broad, and other genomic centers have also setup *seqr* instances (extended and otherwise) at their locations to analyze data. At Broad CMG, *matchbox*, *seqr*, and their ecosystem of support applications, are hosted on the Google cloud platform (<https://cloud.google.com>) and deployed using Kubernetes (<https://kubernetes.io>) and Docker (<https://www.docker.com>). *seqr* groups data into projects, families and individuals. Typically an individual is visible to everyone who has access to the project to which the individual belongs. One or more projects originate from a collaborator. A project team typically consists of a collaborator and a lead analyst from the Broad CMG. At the time of writing, to add a patient to the MME via *matchbox* an investigator has to become a collaborator with the Broad's CMG and add their data to *seqr*.

The general workflow as it relates to *seqr* at the Broad's CMG involves gathering variant and pedigree information from collaborators and uploading them to *seqr*, followed by collaborative analysis. Variant information is obtained via raw VCF files that are parsed using the Hail (<https://github.com/hail-is/hail>) framework and inserted into a Elasticsearch (<https://www.elastic.co/products/elasticsearch>) database cluster. The *seqr* web application uses this distributed database as a variant store along with an instance of a PostgreSQL database (<https://www.postgresql.org>) to store operational data such as usernames, project structures and other tables required by the applied Django framework (<https://www.djangoproject.com>).

Having *seqr* be the frontend to *matchbox* has quite a few benefits. First, it functions as a user interface and provides a one-click process to share patient information with the MME. *seqr* fully centralizes and integrates all information that aids in moving a patient towards a genetic diagnosis. Another benefit of being embedded and proxied through *seqr* is that *matchbox* is not exposed directly to the internet. For an extra layer of security, only a single port is exposed to allow access from *seqr*. *seqr* and *matchbox* are distinct entities and only data from a subset of the over 17,000 patients in *seqr* are submitted to MME. Patients for whom causal candidate genes are identified and the data owner's permission is gained are inserted into *matchbox*.

The family page is the workhorse of the *seqr* system and gives the ability to centralize phenotypes, genotypes, clinician analyst notes, tag variants, and at one glance give a summary of progress on a patient or family. We have embedded the PhenoTips (Girdea et

al., 2013) application within *seqr* to help gather phenotype data efficiently (Figure 1). Once the project team determines a causal candidate gene, they can easily share it in the MME, along with phenotype information via *matchbox*. After the genotype and/or phenotype information is inserted into *matchbox*, it is visible to, and automatically searched within, the MME. Search results are summarized in the family page and updated at every visit. For a detailed view and record keeping functions, or to update their submission, users can visit the detailed search page (Figure 2, Figure 3).

Once shared, match information can be viewed via the *seqr* family page. Analysts can easily identify any new matches that have arisen since their last visit. Two types of matching can be described: “passive matching” that happens once data is shared in *matchbox* and the submitter is contacted by an external party, who is now able match against the submission, or “proactive”, whereby *seqr*-user views matches and reaches out to the other party. We monitor incoming contacts from such “passive” matches for prompt response and support; automated alerts also serve as notification of searches being performed and matches made.

Downloading the source code and starting an instance

We encourage usage and modification of the *matchbox* code as needed and firmly believe that it must be a community driven application. *matchbox* is currently developed and maintained by the Broad CMG and the Monarch Initiative. *matchbox* source code is open source and can be found at <https://github.com/macarthur-lab/matchbox> and carries the BSD clause-3 license (<https://github.com/macarthur-lab/matchbox/blob/master/LICENSE>). *matchbox* must be installed (built) from source code before it can be used. As of writing, we do this via the Docker tool (<https://www.docker.com/>). We provide Kubernetes (<https://kubernetes.io/>) as an advanced installation method to users interested in hosting on the Google cloud (<https://cloud.google.com/>), as is currently done in the Broad CMG.

For installation, a user must have a working MongoDB instance and necessary credentials to connect to it. In the future, we plan to integrate this into Docker as well (for simplicity at the host site, however, that resource is currently kept separate). Docker simplifies the process by reducing the complications that arise from diverse execution environments. It builds and contains the *matchbox* application within a container that can be launched in any Docker supported environment. This installment method is what is recommended and requires a few simple commands to start a fully functional instance of *matchbox*.

matchbox utilization

At the time of writing, *matchbox* carried 257 patients contributing 292 unique candidate genes and 537 unique HPO terms. This is fewer than the number of cases contained in nodes that have been in operation longer such as GeneMatcher (Sobreira, Schiettecatte, Valle, & Hamosh, 2015; Sobreira, Schiettecatte, Boehm, Valle, & Hamosh, 2015) or DECIPHER (Firth et al., 2009) but we hope this approach allows more nodes to join emphasizing federation as a key approach to data sharing. *matchbox* receives about 100 match requests a week and continues to grow with our Center’s gene discovery efforts. A subset of these are match with patients within the *matchbox* system. These matched patients are then

investigated by the project team and data owners of the external patient of the match are contacted for more information and possible collaborations. It must be noted that, consistent with historical practice, only a small subset of these move on to a full collaboration status. Although this is primarily due to candidate genes being entered with limited evidence, additional factors include some of the systematic weakness prevalent in the network, such as non-perfect matching algorithms and minimal input data resulting in imperfect matching.

Discussion

matchbox has been instrumental to the novel gene discovery initiative at the Broad CMG. Due to its integration with *seqr*, our exome and genome analysis interface, novel disease gene candidates that are identified by our analysts are easily entered into the MME using this platform. To date, matches made via *matchbox* have led to three scientific publications (Guissart et al., 2018; Hamdan et al., 2017; Yates et al., 2017) and one additional manuscript has recently been submitted for publication. Five additional publications are in preparation and at least 9 additional genes in active discussion.

The organic growth of the MME in support of matching patients has led to many strengths and flexibilities, but has introduced some inherent challenges to scalability and functionality in its architecture. While the diversity of matching algorithms has allowed exploration of new innovative ideas, it has created some weaknesses; primary among them is the lack of reproducibility of the enumeration of match strength that is generated by the host using its internal algorithms and data. This has made match strength comparison across different centers impossible. If instead, most new sites implement the same system, such as *matchbox*, this diversity will lessen and improve the comparability of results over time. Although this approach does decrease innovation, through ongoing collaboration of a technical working group that supports the MME we continue to work towards improving the tools, algorithms and functions of the MME to ensure continued innovation while working to minimize unnecessary differences.

Developing and supporting the software to join the MME can also be quite intensive and we believe that this has had a detrimental effect on network growth. At its core, the benefits of the MME hinge greatly on the number of patients it contains. Hence, it is crucial that we facilitate continuous growth and add as many patients and nodes into the network as possible. By providing a ready-made, open-source application that can be used almost immediately, new nodes have a stronger incentive to use this tool versus building their own and thus saving the often limited software development resources across sites. Additionally, we foresee *matchbox* as a platform that will allow the plugin of various features as micro-services as it matures allowing continued improvement in the functions of the MME platform.

We see an ecosystem of micro-services that can be started up on host premises that would provide post-processing capabilities as needed, and incrementally address some of the challenges in quality control such as ensuring a sufficient amount of data to be entered to enable more specific matching. We plan to encourage the community to build these with us and by themselves as needed. In the future, we hope nodes that wish to develop their own

matching algorithms would use such microservices to filter or rank *matchbox* results further on their new rule-sets. Microservices are also a means of integrating in technologies such as PhenoPackets (<http://phenopackets.org/>) for data exchange and extending the interpretive capabilities of *matchbox*. One possible architecture could be a microservice functioning downstream of *matchbox* augmenting or filtering the results further before returning them to the querier. In the future, we do intend to provide a nginx-based alternative (<https://www.nginx.com/>) that would function as a proxy for the public API endpoints and hide the sensitive endpoints. We are also considering bundling in a lightweight website as an alternative to the *seqr* for users that require only MME functionality. We plan to work with the community to build this.

Conclusion

We have presented *matchbox* as an open-source tool to facilitate entry into the Matchmaker Exchange. While this application will continue to evolve, we believe that it represents an important contribution to ongoing novel gene discovery initiatives, by further bringing together rare disease patients from around the world for the benefit of patient care and scientific innovation.

Acknowledgements

We would like to thank all members of the Matchmaker Exchange steering committee, technical, and general working groups. We especially like to thank Nara Sobreira, Ada Hamosh, and François Schiettecatte, and Orion J. Buske for their help and guidance when we were designing and launching *matchbox*. We also thank Konrad Karczewski, Matthew Solomonson, Beryl Cummings, Laurent Francioli, and the whole MacArthur-Rehm Rare Disease Group for their help and guidance with algorithm and API design. We would also like to thank members of the Monarch Initiative, especially Peter Robinson, Sebastian Köller, and Chris Mungall for work on the Human Phenotype Ontology and phenotype comparison algorithms. The Broad Center for Mendelian Genomics (UM1 HG008900) is funded by the National Human Genome Research Institute with supplemental funding provided by the National Heart, Lung, and Blood Institute under the Trans-Omics for Precision Medicine (TOPMed) program and the National Eye Institute. This work was also supported by a Broad Ignite award from the Broad Institute. The Monarch Initiative is supported by NIH Office of the Director R24OD011883. Additional support provided by an Australian Government Research Training Program Scholarship, the Australian Genomics Health Alliance & NHMRC (GNT1113531) and the Maurice de Rohan International Scholarship (ABB) and NIH T32GM007748 (MHW) and K12HD052896 (AHOL).

Grant numbers:

National Institutes of Health: T32GM007748, UM1 HG008900, R24OD011883, K12HD052896

Australian Genomics Health Alliance & NHMRC: GNT1113531

References

- Bone WP, Washington NL, Buske OJ, Adams DR, Davis J, Draper D, ... Smedley D (2016). Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, 18(6), 608–617. 10.1038/gim.2015.137 [doi] [PubMed: 26562225]
- Buske OJ, Schiettecatte F, Hutton B, Dumitriu S, Misyura A, Huang L, ... Brudno M (2015). The matchmaker exchange API: Automating patient matching through the exchange of structured phenotypic and genotypic profiles. *Human Mutation*, 36(10), 922–927. 10.1002/humu.22850 [doi] [PubMed: 26255989]

- Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, ... Carter NP (2009). DECIPHER: Database of chromosomal imbalance and phenotype in humans using ensembl resources. *American Journal of Human Genetics*, 84(4), 524–533. 10.1016/j.ajhg.2009.03.010 [doi] [PubMed: 19344873]
- Girdea M, Dumitriu S, Fiume M, Bowdin S, Boycott KM, Chenier S, ... Brudno M (2013). PhenoTips: Patient phenotyping software for clinical and research use. *Human Mutation*, 34(8), 1057–1065. 10.1002/humu.22347 [doi] [PubMed: 23636887]
- Guissart C, Latypova X, Rollier P, Khan TN, Stamberger H, McWalter K, ... Küry S (2018). Dual molecular effects of dominant RORA mutations cause two variants of syndromic intellectual disability with either autism or cerebellar ataxia. *The American Journal of Human Genetics*, 0(0)10.1016/j.ajhg.2018.02.021 Retrieved from [http://www.cell.com/ajhg/abstract/S0002-9297\(18\)30089-2](http://www.cell.com/ajhg/abstract/S0002-9297(18)30089-2)
- Hamdan FF, Myers CT, Cossette P, Lemay P, Spiegelman D, Laporte AD, ... Michaud JL (2017). High rate of recurrent de novo mutations in developmental and epileptic encephalopathies. *American Journal of Human Genetics*, 101(5), 664–685. S0002-9297(17)30377-4 [pii] [PubMed: 29100083]
- Kohler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Ayme S, ... Robinson PN (2017). The human phenotype ontology in 2017. *Nucleic Acids Research*, 45(D1), D876 10.1093/nar/gkw1039 [doi]
- Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, ... Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), 285–291. 10.1038/nature19057 [doi] [PubMed: 27535533]
- Mungall CJ, McMurry JA, Kohler S, Balhoff JP, Borromeo C, Brush M, ... Haendel MA (2017). The monarch initiative: An integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research*, 45(D1), D722 10.1093/nar/gkw1128 [doi]
- Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, ... Rehm HL (2015). The matchmaker exchange: A platform for rare disease gene discovery. *Human Mutation*, 36(10), 915–921. 10.1002/humu.22858 [doi] [PubMed: 26295439]
- Robinson PN, Kohler S, Oellrich A, Sanger Mouse Genetics Project, Wang K, Mungall CJ, ... Smedley D (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research*, 24(2), 340–348. 10.1101/gr.160325.113 [doi] [PubMed: 24162188]
- Smedley D, Jacobsen JO, Jager M, Kohler S, Holtgrewe M, Schubach M, ... Robinson PN (2015). Next-generation diagnostics and disease-gene discovery with the exomiser. *Nature Protocols*, 10(12), 2004–2015. 10.1038/nprot.2015.124 [doi] [PubMed: 26562621]
- Smedley D, Kohler S, Czeschik JC, Amberger J, Bocchini C, Hamosh A, ... Robinson PN (2014). Walking the interactome for candidate prioritization in exome sequencing studies of mendelian diseases. *Bioinformatics (Oxford, England)*, 30(22), 3215–3222. 10.1093/bioinformatics/btu508 [doi]
- Smedley D, Oellrich A, Kohler S, Ruef B, Sanger Mouse Genetics Project, Westerfield M, ... Mungall C (2013). PhenoDigm: Analyzing curated annotations to associate animal models with human diseases. *Database : The Journal of Biological Databases and Curation*, 2013, bat025 10.1093/database/bat025 [doi]
- Smedley D, Schubach M, Jacobsen JOB, Kohler S, Zemojtel T, Spielmann M, ... Robinson PN (2016). A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. *American Journal of Human Genetics*, 99(3), 595–606. S0002-9297(16)30278-6 [pii] [PubMed: 27569544]
- Sobreira NLM, Arachchi H, Buske OJ, Chong JX, Hutton B, Foreman J, ... Matchmaker Exchange Consortium. (2017). Matchmaker exchange. *Current Protocols in Human Genetics*, 95, 9.31.15. 10.1002/cphg.50 [doi]
- Sobreira N, Schiettecatte F, Boehm C, Valle D, & Hamosh A (2015). New tools for mendelian disease gene identification: PhenoDB variant analysis module; and GeneMatcher, a web-based tool for linking investigators with an interest in the same gene. *Human Mutation*, 36(4), 425–431. 10.1002/humu.22769 [doi] [PubMed: 25684268]
- Sobreira N, Schiettecatte F, Valle D, & Hamosh A (2015). GeneMatcher: A matching tool for connecting investigators with an interest in the same gene. *Human Mutation*, 36(10), 928–930. 10.1002/humu.22844 [doi] [PubMed: 26220891]

- Yates TM, Vasudevan PC, Chandler KE, Donnelly DE, Stark Z, Sadedin S, ... Balasubramanian M (2017). De novo mutations in HNRNPU result in a neurodevelopmental syndrome. *American Journal of Medical Genetics. Part A*, 173(11), 3003–3012. 10.1002/ajmg.a.38492 [doi] [PubMed: 28944577]
- Zemojtel T, Kohler S, Mackenroth L, Jager M, Hecht J, Krawitz P, ... Robinson PN (2014). Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Science Translational Medicine*, 6(252), 252ra123 10.1126/scitranslmed.3009262 [doi]

Figure 1.
Family page in *seqr* (using test data for exhibition)

Figure 2:
Family page that summarizes results from MME search (using test data for exhibition)

Figure 3:
Detailed MME search result page (using test data for exhibition)

Table.

All POST methods and the DELETE method expect a JSON data structure.

Protocol	URL	Exposure	Purpose
POST	/match	Public	Matches input JSON patient structure only against patients within <i>matchbox</i> . A MME host has to expose this single endpoint to the internet.
POST	/match/external	Must be protected by a firewall	Queries with the input JSON patient structure only against patients outside of <i>matchbox</i> in the federated MME network. Here <i>matchbox</i> functions as a client that queries the MME network. This endpoint is intended to serve the host institution
POST	/patient/add	Must be protected by a firewall	Adds the input JSON patient structure to <i>matchbox</i> for sharing within MME
GET	/patient/view	Must be protected by a firewall	Returns all patients within <i>matchbox</i>
DELETE	/patient/delete	Must be protected by a firewall	Deletes a patient within <i>matchbox</i> . The JSON payload has the ID of the patient to delete
GET	/metrics/public	Public	Returns patient metrics of <i>matchbox</i> . Only non-sensitive data are shared
GET	/metrics	Must be protected by a firewall	Returns patient metrics of <i>matchbox</i> . Gene level counts and lists are shared