The Male Fertility Gene Atlas – A web tool for collecting and integrating data about epi-/genetic causes of male infertility

Running title: Male Fertility Gene Atlas (MFGA)

H. Krenz[1], J. Gromoll[2], T.Darde[3], F. Chalmel[3], M. Dugas[1], F. Tüttelmann[4]*

[1]Institute of Medical Informatics, University of Münster, 48149 Münster, Germany

[2]Centre of Reproductive Medicine and Andrology, University Hospital Münster, 48149 Münster, Germany

[3]Univ Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) - UMR_S 1085, F-35000 Rennes, France

[4]Institute of Human Genetics, University of Münster, 48149 Münster, Germany

*Correspondence address: Institute of Human Genetics, University of Münster, Vesaliusweg 12-14, 48149 Münster, Germany, frank.tuettelmann@ukmuenster.de

Word count: abstract 276, main text: 3914

Abstract

Interconnecting results of previous OMICs studies is of major importance for identifying novel underlying causes of male infertility. To date, information can be accessed mainly through literature search engines and raw data repositories. However, both have limited capacity in identifying relevant publications based on aggregated research results e.g. genes mentioned in images and supplements. To address this gap, we present the Male Fertility Gene Atlas (MFGA), a web tool that enables standardised representation and search of aggregated result data of scientific publications. An advanced search function is provided for querying research results based on study conditions/phenotypes, meta information and genes returning the exact tables and figures from the publications fitting the search request as well as a list of most frequently investigated genes. As basic prerequisite, a flexible data model that can accommodate and structure a very broad range of meta information, data tables and images was designed and implemented for the system. The first version of the system is published at the URL https://mfga.uni-muenster.de and contains a set of 46 representative publications. Currently, study data for 28 different tissue types, 32 different cell types and 20 conditions is available. Also, ~5,000 distinct genes have been found to be mentioned in at least ten of the publications. As a result, the MFGA is a valuable addition to available tools for research on the epi-/genetics of male infertility. The MFGA enables a more targeted search and interpretation of OMICs data on male infertility and germ cells in the context of relevant publications. Moreover, its capacity for aggregation allows for meta-analyses and data mining with the potential to reveal novel insights into male infertility based on available data.

Keywords: male infertility, genetics, epigenetics, omics, database

Introduction

Male infertility is a prevalent and highly heterogeneous disease for which the underlying causes can currently be identified for only about 30% of male partners in infertile couples (Tüttelmann *et al.*, 2018). In the last years, a growing number of studies have been published on the genetics and epigenetics of male infertility using a broad range of genomic technologies. For an overview see Oud *et al.* (2019). Often, researchers give extensive insight into their findings by providing supplementary data or access to their raw data. Theoretically, this enables clinicians and other researchers to validate those findings or interpret their own results in a broader context; in practice, however, the findability and reusability of this vast set of information is limited. To date, there is no public resource for the field of male infertility that enables clinicians and researchers alike to easily access a comprehensive overview of recent findings, although tools for other diseases have long been established, such as, for example, the Gene ATLAS (Canela-Xandri *et al.*, 2018) that was released in 2017. It provides information on genetic associations of 778 traits identified in genome-wide association studies (Canela-Xandri *et al.*, 2018), male infertility is, unfortunately, not one of them.

Instead of being accessible through a specified tool, research on male infertility relies mainly on general search engines for scientific publications like PubMed or Google Scholar. These engines can be employed to search for information based on keywords of interest, e.g., gene names in combination with specific conditions. However, if the required information can only be found in a supplementary table or figure, these search engines are unable to mark the corresponding publication as relevant. Other sources of information are raw data repositories like Gene Expression Omnibus (GEO) which provides access to the raw data files of microarray and genomic data from many publications (Barrett *et al.*, 2013), or Sequence Read Archive (SRA). This is a marked achievement for the scientific community, but such repositories do not allow users to find a data file based on a gene or variant of interest. For the most part, publications of interest have to be identified in advance, and comprehensible insight into the data can only be achieved by reanalysing it, using complex bioinformatics pipelines.

To address this need specifically for the reproductive sciences, Chalmel and colleagues established the ReproGenomics Viewer (RGV) in 2015 (Darde *et al.*, 2019; Darde *et al.*, 2015). It provides a valuable resource of manually-curated transcriptome and epigenome data sets processed with a standardised pipeline. RGV enables the visualisation of multiple data sets in an interactive online genomics viewer and, thus, allows for comparisons across publications, technologies and species (Darde *et al.*, 2019). However, the RGV is not designed to show downstream analysis results such as differentially expressed genes or genomic variation. Also, it is not possible to identify relevant publications based on genes of interest. Other tools in this field are GermOnline (Lardenois *et al.*, 2010) and SpermatogenesisOnline (Zhang *et al.*, 2013). However, both have not been maintained in recent years.

In order to offer a public platform that provides access to a comprehensive overview of research results in the field of epi-/genetics of male infertility and germ cells and to bridge the gap between textual information in publications on the one hand and complex data sets on the other hand, we designed, developed, and now introduce the publicly available Male Fertility Gene Atlas (MFGA, https://mfga.uni-muenster.de). Its objective is to provide fast, simple and straightforward access to aggregated analysis results of relevant publications, namely by answering questions like “What is known about the gene *STAG3* in the context of male infertility?” or “Which genes have been identified to be associated with azoospermia?“. To this end, we created an advanced search interface as well as comprehensive overviews and visualisations of the publications and search results. A basic prerequisite was that we had to design and implement a data model that can accommodate and structure a very broad range of meta-information, data tables and images.

Materials and methods

Requirements engineering

The MFGA is designed to support researchers and clinicians in the field of male infertility in finding and evaluating the analysis results of relevant publications. To ensure that the MFGA

offers a relevant scope and focusses on the central needs and interests of the targeted user group, an extensive requirements engineering process had to be employed. This process aimed at defining the most relevant parts of the system design; in case of the MFGA, this means answering the following five questions:

1. What kind of data should be available?
2. How should the database be maintained and kept up to date?
3. Who should be authorised to access the data on the website?
4. What interface should be provided to access the data?
5. How should the data be visualised?

Since user-centric design is proven to be a critical factor of success for software projects (Maguire and Bevan, 2002), the requirements were acquired in close cooperation with a large group of prospective users of the MFGA. These researchers and clinicians are part of the Münster-based clinical research unit (CRU326) (http://www.male-germ-cells.de) and provided input on a broad range of aspects of male infertility, germ cell development as well as sperm morphology, motility, and function. The development of the MFGA is based on the software development and requirements engineering paradigm Rapid Prototyping, which focusses on developing early testing versions of the product to enable informed user feedback and iterative improvements (Budde *et al.*, 1992; Gordon and Bieman, 1995).

The requirements analysis was based on a set of 39 representative and relevant publications/data sets that the MFGA should be able to host, provided by members of the CRU326. The publications were highly heterogeneous and covered a broad range of methodologies, from GWA studies to targeted genotyping, microarray, bulk or single-cell RNA sequencing as well as methylation. The complete list is available in Suppl. Tab. S1 and has already been included into the MFGA database.

Current state of requirements

The requirements regarding the kind of data that should be available in the MFGA can be summarised in two points: First, the MFGA should host data representing the analysis results of a comprehensive set of publications on male infertility. Generally speaking, the MFGA needs to be able to deal with publications of arbitrary form and content. To account for the extremely high degree of heterogeneity between publications, a very flexible, modular data model is required. However, to enable the MFGA's central service of offering selective data retrieval operations, it is crucial to identify, standardise and tag recurrent relevant information items and to curate them manually for each publication. Whenever applicable, standardisation should be based on ontologies as provided by OBO Foundry (Smith *et al.*, 2007). Second, the MFGA should provide complementary data from external databases to increase the comprehensibility of information, e.g., to explain the background of gene names and provide links to more details. However, the MFGA is not supposed to contain any raw data; instead the system should link to the corresponding repositories if provided by a publication. Also, the MFGA is not planned for processing raw data by itself; the system is only intended to show aggregated result data provided by publications.

The database of the MFGA should be fully public and accessible to all researchers and clinicians in the field of male infertility. Maintenance and curation of selected publications should be provided by the development team. In order to keep the database up to date in a structured way, a recurring process for preparing new relevant publications will be implemented (Fig. 1). It comprises three main steps: selection of relevant publications, manual curation according to the requirements of the MFGA database and data upload. Prospectively, it should be triggered once per month and combine the proposed publications of four different sources into one prioritised list. The four information sources are (1) the list of not yet processed publications from the previous month, (2) newly published manuscripts from the CRU326, (3) other relevant new publications identified by the MFGA team, and, as an additional unbiased source, (4) recent publications queried from PubMed (query: "(male OR men) AND (fertility OR infertility) AND (gene OR genetic) AND ("YYYY/MM/01"[Date -

Publication] : "YEAR/MM/31"[Date - Publication])"). Publications will be curated and uploaded by priority. Eventually, users are supposed to be enabled to act as registered contributors and support selection, curation and upload of content. Thus, a user management and authentication system is required.

Regarding the interface for accessing the information, users should be able to view a complete list of publications and their key features to get a quick overview on the full content of the data base. Additionally, an advanced search function should provide identifying and showing subsets of relevant publications based on the occurrence of gene names or IDs in tables and figures as well as meta-information like data type, OMICS, processed tissues/cells and conditions/species. Detailed information and results of each individual publication should be displayed by comprehensible, standardised overviews, including images, tables and on-demand plots. As a special feature, the MFGA should display data from different publications in an aggregated way to allow for meta-analyses and integrated analyses. All pages should be enriched with appropriate additional information from external databases and with cross references, allowing for quick navigation through the atlas.

Results

IT architecture

The MFGA has been implemented as a modern *Java* web application that is securely hosted on a server at the University of Münster, Germany, and can be accessed freely via the URL https://mfga.uni-muenster.de. Its architecture (Suppl. Fig. S1) is based on the specific requirements, identified during requirements engineering and explained in detail in the supplement. For a good user experience, the graphical user interface utilises the *Bootstrap* library which enables a smooth adjustment of the application to different browsers and devices (Otto *et al.*, 2020) and familiar design elements. On the server side, Spring Security framework is employed to provide secure authentication processes and manage data access privileges (Pivotal Software Inc., 2019).

Data model

The specific relational data model of the MFGA as well as the preparation process for publication data are a direct result of the requirements analysis and were developed based on the initial set of relevant publications. They are the main prerequisite for enabling complex search queries, aggregation functions and meta analyses on the MFGA's database. For representing the highly heterogeneous publications, a flexible and thus modular data model was designed (Suppl. Fig. S2). It is based on information items that were identified to be preserved over large subsets of publications on male infertility and standardised to seven classes of information items: *publication meta information*, *data set meta information*, *processed tissue type*, *processed cell type*, *cohort's condition*, *image* and *table* (Tab. 1). For each publication, they can freely be combined in any quantity including zero. In order to prevent inconsistencies in the database caused by misspelling or synonymous terms, structured data entry with drop-down lists and ontologies is implemented. Currently, the MFGA employs BRENDA Tissue Ontology (BTO) (Gremse *et al.*, 2011) for tissue type definition, Cell Ontology (CL) (Diehl *et al.*, 2016) for cell types and Human Phenotype Ontology (HPO) (Köhler *et al.*, 2019) for conditions. Further technical specification, especially representation and annotation of data tables and images, is explained in the extended methods (see supplement).

Available content

Currently there are 46 publications available on the MFGA, and they comprise information on 101 data sets (Fig. 2), 28 different tissue types, 32 different cell types and 20 conditions. The majority of data sets contains data on the transcriptome (22 single-cell RNA sequencing, 12 bulk RNA sequencing & 3 microarray). The second largest group of data sets contains information on the genome / exome (13 targeted genotyping, 9 genome-wide association studies, 4 whole-genome sequencing & 4 sanger sequencing). Additionally, 6 data sets on whole-genome bisulfite sequencing and 2 on targeted deep bisulfite sequencing are available.

The most frequent conditions in the data sets are variants of abnormal spermatogenesis, e.g. azoospermia and oligozoospermia. Testicular tissue has been processed by 32 proteome, methylome and transcriptome data sets. The predominantly represented cell types are embryonic stem cells, primordial germ cells, spermatogonia, spermatocytes, and spermatids. Also, ~5,000 distinct genes have been found to be mentioned in at least ten of the publications, with the top genes being *TEX11*, *TEX14*, *TEX15*, *SYCP3*, *SMC1B*, *REC8* and *HORMAD1*. The full list of 46 publications can be accessed via the URL https://mfga.uni-muenster.de/publication.html.

Functionality

The MFGA provides a web interface for fast, simple, and straightforward access to the results of publications on the epi-/genetics of male infertility and germ cells. For this purpose, an advanced search form is provided on the *Search* tab of the atlas (Suppl. Fig. S3), enabling the specification of search requests for relevant data sets. This function supports search terms from seven categories: OMICS, data type (e.g., single-cell RNA sequencing or targeted sequencing), condition, species, cell type, tissue type and gene/ID. For each of the first six categories, one can choose from a list of search terms equivalent to the information items stored in the database. Condition, cell type and tissue type are based on appropriate ontologies: HPO (Köhler *et al.*, 2019), CL (Diehl *et al.*, 2016) and BTO (Gremse *et al.*, 2011). The number of available data sets is shown in brackets after each search term. Gene names and IDs can be specified in a text field as a comma-separated list. Also, there are three search modalities: (1) users can perform a broad search to identify all data sets that are annotated with at least one of the specified search terms, (2) users can perform a more targeted search for all data sets that are annotated with at least one of the specified search terms per category and (3) users can perform a very specific search for all data sets that are annotated with each of the search terms (default option). As an example, Fig. 3 shows an extract of the output of a simple search request for data sets in the MFGA database containing information on the gene *STAG3* (Suppl. Fig S2 for the full screenshot). This refers back to the introductory question:

"What is known about the gene *STAG3* in the context of male infertility?". For the following examples, the gene *STAG3*, which is located on chromosome 7, encodes a protein involved in meiosis and can be related to male infertility (van der Bijl *et al.*, 2019), is employed to explain the general functionalities of the MFGA. A detailed *Walk Through* is provided on https://mfga.uni-muenster.de/walkThrough.html.

Executing a search request on the MFGA results in a list of data sets matching the requirements, represented via charts and result tables (Fig. 3 and Suppl. Fig. S3). The left chart shows the number of returned data sets grouped by publication. Multiple data sets matching the search request in one publication can indicate that the authors provided further proof for their findings, such as e.g., van der Bijl *et al.* (2019), who screened two cohorts of infertile men. A textual summary of the data sets returned by the search request and some important meta-information is given by the *Data sets* table. All tables in the MFGA format are fully interactive such that the columns can be sorted, filtered and plotted online. Whenever the search contains genes or IDs, the right chart represents their frequencies in the different publications. This provides an initial estimation of relevance. Additionally, a second table lists their specific occurrences in data tables and figures. In the case when a search request is targeted at revealing relevant genes, such as, e.g., in the introductory question: "Which genes have been identified to be associated with azoospermia?" (Suppl. Fig. S4), the right chart and second table present the genes that occur most frequently in data tables of the returned data sets.

The search results returned by the MFGA are designed for quick navigation and information access. Therefore, many cross references are provided: Table entries directly link to overview pages of the corresponding publication and its data sets. Data tables and images are linked in the MFGA as well. However, this functionality is restricted to publications that are published under an open access license. As an example, Fig. 3 shows that the publication van der Bijl *et al.* (2019) mentions *STAG3*. Clicking on the data set title leads to the overview page of that publication (Suppl. Fig. S5). Also, data tables containing the searched gene can directly be accessed in MFGA by a single click, e.g., supplementary table 3 of van der Bijl *et al.* (2019)

which is shown partly in Fig. 4 (full version in Suppl. Fig. S6). For a deeper analysis of the underlying read data of individual data sets, the MFGA provides a link to the RGV, whenever a data set is available in both tools, such as, e.g., Hammoud *et al.* (2015) in Fig. 3.

Data from external public databases has been integrated in order to enrich information on the publications shown in the MFGA. Throughout the application, gene names can be selected in table cells and plots to open up an overlay with further explanations (for an example, see Fig. 5). The overlay bundles textual information from RefSeq (O'Leary *et al.*, 2016) and HGNC (Yates *et al.*, 2017; HGNC Database, 2018). Additionally, a link to the corresponding GeneCards page (Stelzer *et al.*, 2016) is provided, as well as a shortcut to the MFGA search for that gene. Data from the GTEx project (Lonsdale *et al.*, 2013; The Broad Institute of MIT and Harvard, 2019) is employed to show the gene's expression in different tissues scaled by the largest median. The expression of the gene in testis tissue is highlighted in red.

Discussion

The MFGA provides a public platform to support researchers and clinicians in obtaining an overview of the recent findings in the field of male infertility and germ cells. The web-based tool includes recent libraries and methods for an improved user experience and straightforward usability. In order to enable an advanced search for relevant publications, a relational data model has been developed and implemented for knowledge representation. It records the recurrent information items in a structured and consistent way and can accommodate arbitrary publications from the field of male infertility, regardless of the kind of data analysis and technology. The search form enables a broad range of simple to very complex search queries such as "What is known about the gene *STAG3* in the context of male infertility?", "Which genes have been identified to be associated with azoospermia?" or "Which genes have been found to be expressed in Sertoli cells of human testis tissue using single cell RNA sequencing?" (Suppl. Fig. S7).

The main purpose of the MFGA is to enable researchers and clinicians to consider their own research results in the context of other relevant literature. To this end, the system enables a highly selective search for studies with comparable conditions and parameters and offers the means to quickly review their main analysis results. Additionally, searches for genes can be performed in order to identify data sets containing the corresponding genes in their analysis results. This functionality supports the validation of candidate genes. Further, supplementary information from various sources is embedded into the MFGA, e.g., RefSeq (O'Leary *et al.*, 2016), HGNC (Yates *et al.*, 2017; HGNC Database, 2018) and GTEx project (Lonsdale *et al.*, 2013; The Broad Institute of MIT and Harvard, 2019).

Compared to general search engines like Google Scholar and PubMed, with millions of records, the MFGA will always return smaller numbers of search results. However, in the domain of male infertility, the relevance of the individual results in relation to the corresponding search terms used in the MFGA is expected to be greater than when using those same search terms in a large search engine. Search engines for literature are usually restricted to mining textual information of publications and cannot consider information that is presented in tables, images, or supplementary material. The MFGA, however, is specifically designed for that task, which is enabled by its key features: A standardised data representation and disease-specific manual curation.

Another problem occurs when relevant data sets are searched in raw data repositories like GEO or processed data resources like RGV (Darde *et al.*, 2015; Darde *et al.*, 2019). While the information that these tools are able to provide is much more detailed than overviews in the MFGA, they cannot be used to identify a data set based on, for example, a list of genes that are supposed to be differentially expressed, since raw data does usually not include that kind of annotation. The MFGA facilitates searching through aggregated result data and, thus, identifying relevant data sets. Once identified, such data sets might then be further investigated in the RGV (Darde *et al.*, 2015; Darde *et al.*, 2019) or by processing GEO data with bioinformatics tools. Since RGV (Darde *et al.*, 2015; Darde *et al.*, 2019) and MFGA

complement each other in their functionality, the MFGA links to RGV whenever a data set is present in both tools; further, an even closer integration is planned.

Regarding tools that might appear similar to the MFGA, Zhang *et al.* (2013) provide a tool termed SpermatogenesisOnline for searching individual genes in the context of spermatogenesis based on a set of publications from 2012 and earlier. However, it remains unclear which publications were included and whether or not they have been updated since then. Another tool from Lardenois *et al.* (2010), GermOnline, provides functional information about individual genes and access to transcriptomics data from microarrays on germline development from 2008. Both tools have, to our knowledge, not been updated and are, thus, not very useful anymore.

There are some limitations to the approach of the MFGA. Since analysis results are not reproduced using in-house pipelines, the MFGA is restricted to the analysis results authors are presenting in their publications. In the case where a publication provides, e.g., only significant SNPs from a GWA study, the MFGA, too, reports only these. The only meta-analyses enabled are those that use the aggregation of information in the MFGA, e.g., the top score of gene appearances, to approximate gene importance; the information cannot indicate correlation or even causality. Thus, the MFGA focusses on proposing plausible research hypotheses. Finally, full functionality of the MFGA can only be provided for publications under an open access license. Publications that allow no or restricted reuse are only partly integrated.

Prospectively, the MFGA will be updated and enlarged based on the proposed content management process (Fig. 1). In this context, a web form will be implemented to enable users to propose publications that are, from their point of view, still missing. Additionally, a closer integration of RGV and MFGA is planned, e.g. by coupling the publication submission forms, as well as further tools for meta-analyses.

In conclusion, the MFGA is a valuable addition to available tools for research on the epi-/genetics of male infertility. It helps fill the gap between pure literature searches as provided by PubMed and raw data repositories like GEO or SRA. The MFGA enables a more targeted

search and interpretation of OMICS data on male infertility and germ cells in the context of relevant publications. Moreover, its capacity for aggregation allows for meta-analyses and data mining with the potential to reveal novel insights into male infertility based on available data. Ultimately, by combining RGV and MFGA with AI methods, we aim to develop a powerful gene prioritisation system dedicated to male infertility similar to the GPSy tool (Britto *et al.*, 2012).

Acknowledgements

We are grateful to all members of the Clinical Research Unit (CRU) 'Male Germ Cells' who contributed through either identifying relevant publications, supporting publication curation or providing feedback during the development of the MFGA: Michael Storck, Marius Wöste, Nina Neuhaus, Sven Berres, Sandra Laurentino, Lina Franziska Lanuza Pérez, Corinna Friedrich, Maria Schubert, Nikithomas Loges, Isabella Aprea, Jana Emich, Eva Maria Mall, Johanna Raidt, Nadja Rotte, Alexander Busch, Sara Kim Plutta, Jascha Henseler, Anna Natrup. We thank Dr. Celeste Brennecka for language editing of the manuscript.

Authors' roles

H.K. designed and developed the MFGA data model and system architecture, implemented the tool and drafted the manuscript. J.G. provided critical input during the development of the MFGA and first drafts of the manuscript. T.D. and F.C. contributed to integrating the mutual links between RGV and MFGA. M.D. and F.T. contributed to the system design and supervised the whole study. All authors critically revised the manuscript and approved the final version.

Funding

This work was carried out within the frame of the German Research Foundation (DFG) Clinical Research Unit 'Male Germ Cells: from Genes to Function' (CRU326).

Conflict of interest

The authors declare no conflicts of interest.

References

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic acids research* 2013;**41**:D991-D995.

Britto R, Sallou O, Collin O, Michaux G, Primig M, Chalmel F. GPSy: a cross-species gene prioritization system for conserved biological processes--application in male gamete development. *Nucleic acids research* 2012;**40**:W458-65.

Budde R, Kautz K, Kuhlenkamp K, Züllighoven H. What is prototyping? *Info Technology & People* 1992;**6**:89–95.

Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. *Nature genetics* 2018;**50**:1593–1599.

Chinosi M, Trombetta A. BPMN: An introduction to the standard. *Computer Standards & Interfaces* 2012;**34**:124–134.

Darde TA, Lecluze E, Lardenois A, Stévant I, Alary N, Tüttelmann F, Collin O, Nef S, Jégou B, Rolland AD, et al. The ReproGenomics Viewer: a multi-omics and cross-species resource compatible with single-cell studies for the reproductive science community. *Bioinformatics (Oxford, England)* 2019;**35**:3133–3139.

Darde TA, Sallou O, Becker E, Evrard B, Monjeaud C, Le Bras Y, Jégou B, Collin O, Rolland AD, Chalmel F. The ReproGenomics Viewer: an integrative cross-species toolbox for the reproductive science community. *Nucleic acids research* 2015;**43**:W109-W116.

Diehl AD, Meehan TF, Bradford YM, Brush MH, Dahdul WM, Dougall DS, He Y, Osumi-Sutherland D, Ruttenberg A, Sarntivijai S, et al. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of biomedical semantics* 2016;**7**:44.

Gordon VS, Bieman JM. Rapid prototyping: lessons learned. *IEEE Softw.* 1995;**12**:85–95.

Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, Schomburg D. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic acids research* 2011;**39**:D507-13.

Hammoud SS, Low DHP, Yi C, Lee CL, Oatley JM, Payne CJ, Carrell DT, Guccione E, Cairns BR. Transcription and imprinting dynamics in developing postnatal male germline stem cells. *Genes & development* 2015;**29**:2312–2324.

HGNC Database. *HGNC Database: retrieved in November 2018*: HUGO Gene Nomenclature Committee (HGNC), European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom, 2018.

Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine J-P, Gargano M, Harris NL, Matentzoglu N, McMurry JA, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic acids research* 2019;**47**:D1018-D1027.

Lardenois A, Gattiker A, Collin O, Chalmel F, Primig M. GermOnline 4.0 is a genomics gateway for germline development, meiosis and the mitotic cell cycle. *Database the journal of biological databases and curation* 2010;**2010**:baq030.

Lonsdale J, Thomas J, Salvatore Mea. The Genotype-Tissue Expression (GTEx) project. *Nature genetics* 2013;**45**:580–585.

Maguire M, Bevan N. User Requirements Analysis. In: Hammond J, Gross T, Wesson J (eds). *Usability*. Boston, MA: Springer US, 2002.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* 2016;**44**:D733-D745.

Otto M, Thornton J, contributors aB. *Introduction*. Retrieved 09 January 2020. from https://getbootstrap.com/docs/4.4/getting-started/introduction/, 2020.

Oud MS, Volozonoka L, Smits RM, Vissers LELM, Ramos L, Veltman JA. A systematic review and standardized clinical validity assessment of male infertility genes. *Human reproduction (Oxford, England)* 2019;**34**:932–941.

Pivotal Software Inc. *Spring Security*. Retrieved 10 December 2019. from https://spring.io/projects/spring-security, 2019.

R Development Core Team. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2008.

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;**25**:1251–1255.

Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Stein TI, Nudel R, Lieder I, Mazor Y, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current protocols in bioinformatics* 2016;**54**:1.30.1-1.30.33.

The Broad Institute of MIT and Harvard. *GTEx Portal. File: GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_median_tpm.gct.gz*. Retrieved 09 December 2019. from https://gtexportal.org/home/datasets, 2019.

Tüttelmann F, Ruckert C, Röpke A. Disorders of spermatogenesis: Perspectives for novel genetic diagnostics after 20 years of unchanged routine. *Medizinische Genetik Mitteilungsblatt des Berufsverbandes Medizinische Genetik e.V* 2018;**30**:12–20.

van der Bijl N, Röpke A, Biswas U, Wöste M, Jessberger R, Kliesch S, Friedrich C, Tüttelmann F. Mutations in the stromal antigen 3 (STAG3) gene cause male infertility due to meiotic arrest. *Human reproduction (Oxford, England)* 2019;**34**:2112–2119.

Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA. Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic acids research* 2017;**45**:D619-D625.

Zhang Y, Zhong L, Xu B, Yang Y, Ban R, Zhu J, Cooke HJ, Hao Q, Shi Q. SpermatogenesisOnline 1.0: a resource for spermatogenesis based on manual literature curation and genome-wide data mining. *Nucleic acids research* 2013;**41**:D1055-62.

449 **Table 1. Information classes and items available for each publication.**

Information Class	Information items[1]	
Publication meta information	title	publishing date
	author	important genes*
	citation	link to publication
	abstract	link to repository*
Data set meta information	title*	OMICS type
	technology*	reference genome*
	data type	species
Processed tissue types	name	maturity*
	description*	potential*
	no of subjects*	species
	no of probes per subject*	reference genome*
Processed cell types	name	maturity*
	description*	potential*
	no of subjects*	species
	no of cells per subject*	reference genome*
Cohort's conditions	Human phenotype ontology* id (Köhler *et al.*, 2019)	name
	size of cohort*	comment*
Images	title	description*
	annotation of genes or relevant IDs	path
Tables	title	description*
	annotation of standard columns (gene names, relevant IDs, loci, …)	annotation of all additional columns as numeric or text

450

451 [1]*Optional items are marked with *.*

452

Figure 1. The process of identifying, selecting and prioritising novel publications depicted based on Business Process Model and Notation (Chinosi and Trombetta, 2012). The process is triggered monthly. Potential publications are collected and merged into a list. This list is then prioritised by the MFGA team. Subsequently, curation and upload are performed in the determined order.

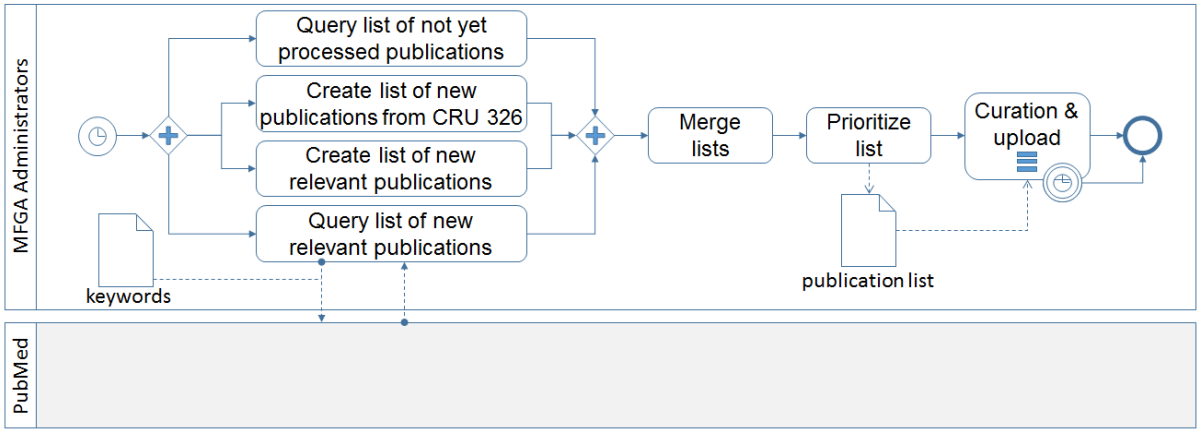

Figure 2. Data sets available on the MFGA grouped by technology. Currently 101 data sets corresponding to 46 publications are fully curated and publically accessible. Abbreviations: seq – sequencing, GWAS – genome-wide association study, WGBS – whole genome bisulfite sequencing, BS – bisulfite. Plot created with R (R Development Core Team, 2008).

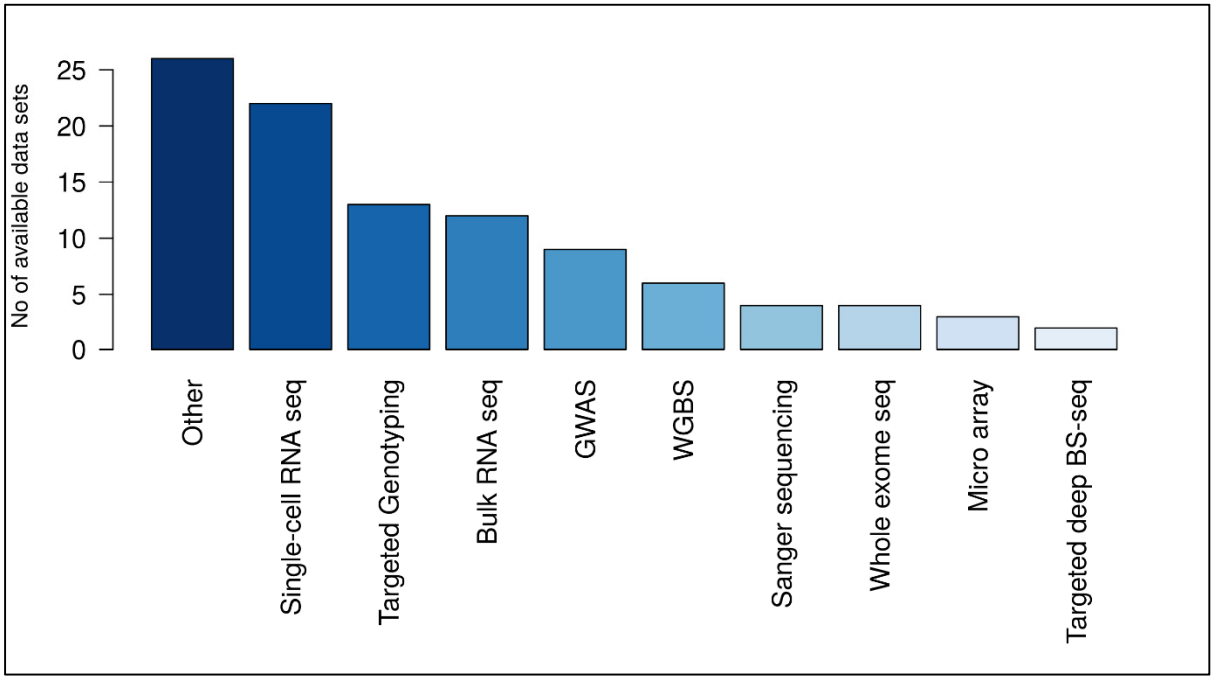

Figure 3. Example screenshot of the MFGA's search functionality (https://mfga.uni-muenster.de/search.html). It shows part of the result of the search for the gene *STAG3* in the MFGA database. The side bar on the left contains the search form. Here, *STAG3* is typed into the search field and search option 'all search terms' is checked. In the main window, plots show the number of returned data sets and the frequency of *STAG3* in data tables of the corresponding publications. Below, the returned data sets are presented in a table enriched with further meta-information, linking to the overview of data set and publication. The second table shows the specific data tables and images in which the gene was found. Here, tables are cropped and screenshot is compressed for better representability. See Suppl. Fig. S3 for original screenshot. Retrieved 10 January 2020.

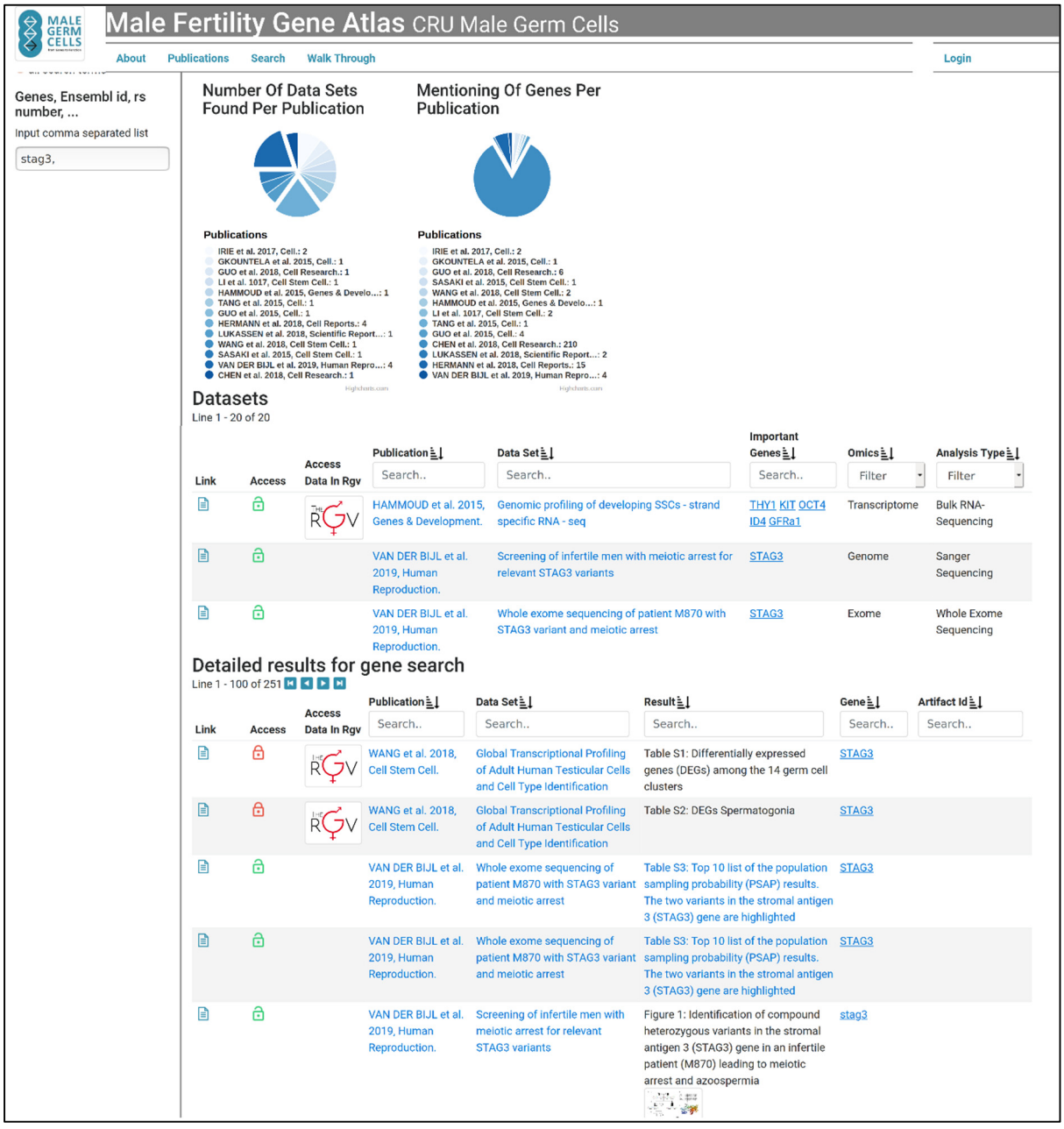

Figure 4. Example screenshot of the MFGA's functionality to represent data from publications (https://mfga.uni-muenster.de/publication/results?ssDetailsId=3968811). It shows part of supplementary table 3 of van der Bijl *et al.* (2019) in MFGA format. The table is represented in an interactive way and can directly be filtered and sorted. Additionally, histograms can be plotted for the table columns, here, e.g., based on column "Testis Expression". For the full screenshot see Suppl. Fig. S6. Retrieved 10 January 2020.

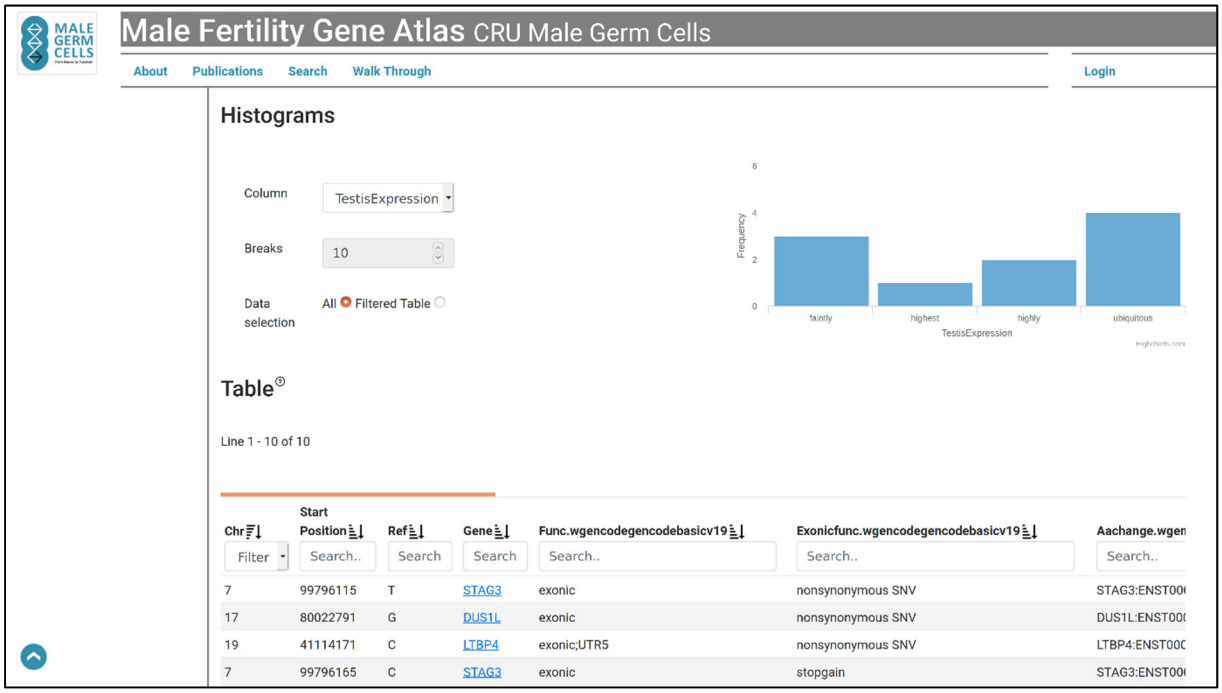

Figure 5. Screenshot of the MFGA showing the overlay for additional information on genes (https://mfga.uni-muenster.de/publication/results?ssDetailsId=3968811). Here, as an example, the overlay for the gene STAG3 is shown. The textual information originates from RefSeq (O'Leary *et al.*, 2016) and HGNC (Yates *et al.*, 2017). Additionally, gene expression in different tissues is shown based on the data from the GTEx project (Lonsdale *et al.*, 2013). Links for searching the gene in the MFGA database and opening the corresponding GeneCards entry (Stelzer *et al.*, 2016) are provided. Retrieved 10 January 2020.