The Male Fertility Gene Atlas: a web tool for collecting and integrating OMICS data in the context of male infertility

Henrike Krenz, Jörg Gromoll, Thomas A Darde, Frédéric Chalmel, Martin Dugas, Frank Tüttelmann

The Male Fertility Gene Atlas – A web tool for collecting and integrating OMICS data in the context 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 446, main text: 3987 (incl. references)

Abstract

Study question: How can one design and implement a system that provides a comprehensive overview of research results in the field of epi-/genetics of male infertility and germ cells?

Summary answer: Working at the interface of literature search engines and raw data repositories, the newly developed Male Fertility Gene Atlas (MFGA) provides a system that can represent aggregated results from scientific publications in a standardised way and perform advanced searches based on e.g. the conditions (phenotypes) and genes related to male infertility.

What is known already: PubMed and Google Scholar are established search engines for research literature. Additionally, repositories like Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) provide access to raw data. Selected processed data can be accessed by visualisation tools like the ReproGenomics Viewer (RGV).

Study design, size, duration: The MFGA was developed in a time frame of 18 months under a rapid prototyping approach.

Participants/materials, setting, methods: In the context of the Clinical Research Unit 'Male Germ Cells' (CRU326), a group of around 50 domain experts in the fields of male infertility and germ cells helped develop the requirements engineering and feedback loops. They provided a set of 39 representative and heterogeneous publications to establish a basis for the system requirements.

Main results and the role of chance: The MFGA is freely available online at https://mfga.uni-muenster.de. To date, it contains 115 data sets corresponding to 54 manually-curated publications and provides an advanced search function based on study conditions, meta-information and genes, whereby it returns the publications' exact tables and figures that fit the search request as well as a list of the most frequently investigated genes in the result set. Currently, study data for 31 different tissue types, 32 different cell types and 20

conditions is available. Also, ~8,000 and ~1,000 distinct genes have been found to be mentioned in at least ten and fifteen of the publications, respectively.

Large scale data: Not applicable because no novel data was produced.

Limitations, reasons for caution: For the most part, the content of the system currently includes the selected publications from the development process. However, a structured process for the prospective literature search and inclusion into the MFGA has been defined and is currently implemented.

Wider implications of the findings: The technical implementation of the MFGA allows for accommodating a wide range of heterogeneous data from aggregated research results. This implementation can be transferred to other diseases to establish comparable systems and generally support research in the medical field.

Study funding/competing interest(s): 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). The authors declare no conflicts of interest.

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

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. In order to achieve a good acceptance and readiness to use the system among users, an extensive requirements engineering process was employed prior to implementation. It was based on a set of 39 highly heterogeneous publications/data sets that are relevant for and representative of research in male infertility and germ cells (see Suppl. Tab. S1) as well as on constant feedback of a large group of researchers and clinicians in the field of male infertility and germ cells.

The requirements identified in this process can be summarized as follows: First, the MFGA should host data representing the analysis results of a comprehensive set of publications on male infertility, including information on publication meta information, data set meta information, processed tissue types, processed cell types, cohort's conditions/species, images and tables (Tab. 1). Second, this data should be enriched and complemented by data from external databases, e.g. to provide general information on genes and enable standardisation of terms with ontologies. Third, data sets should be searchable and identifiable based on the occurrence of gene names or IDs in tables and figures of publications as well as on meta-information. Forth, maintenance and updating of the platform should be provided. Last, all data in the MFGA should be publically available.

For technical details, please see sections Requirements engineering, Current state of requirements, IT architecture and Data model in the Extended Methods of the supplement, as well as Suppl. Fig. S1, S2 and S3.

Results

Available content

To date (30 April 2020) there are 54 publications available on the MFGA, and they comprise information on 115 data sets (Fig. 1), 31 different tissue types, 32 different cell types and 20

conditions. The majority of data sets contains data on the transcriptome (23 single-cell RNA sequencing, 12 bulk RNA sequencing & 11 with other techniques). The second largest group of data sets contains information on the genome / exome (13 targeted genotyping, 9 genome-wide association studies & 15 with other techniques). Additionally, 10 data sets on methylome and 10 on proteome are available as well as 12 uncategorized data sets.

The most frequent conditions in the data sets are variants of abnormal spermatogenesis, e.g. azoospermia and oligozoospermia. Testicular tissue has been processed for 36 data sets. The predominantly represented cell types are spermatogonia, sperm cells, embryonic stem cells, spermatocytes and primordial germ cells. Also, ~8,000 and ~1,000 distinct genes have been found to be mentioned in at least ten and fifteen of the publications, respectively, with the top genes being *TEX11*, *TEX14*, *DNMT1*, *SMC1B*, *TEX15*, *GGH*, *GLUL*, *HORMAD1*, *STK31*, *SYCP3* and *TOP2A*. The full list of 54 publications is available in Suppl. Tab. 1 and 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. S4), 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 available search terms. Condition, cell type and tissue type are based on appropriate ontologies: Human Phenotype Ontology (Köhler *et al.*, 2019), Cell Ontology (Diehl *et al.*, 2016) and BRENDA Tissue Ontology (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. Two checkboxes enable including synonyms of the entered gene names and IDs into the search. 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. 2 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. S4a and Suppl. Fig. S4b 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. 2 and Suppl. Fig. S4a and Suppl. Fig. 4b). 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 and plots in the MFGA format are fully interactive. Tables can be sorted, filtered and plotted online and plots provide e.g. more information on mouse over. 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. S5a and

Suppl. Fig. 5b), 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. 2 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. S6). 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. 3 (full version in Suppl. Fig. S7). 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., Irie *et al.* (2015) in Fig. 2.

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. 4). 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 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. S8a and Suppl. Fig. 8b).

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 specifically. 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 (Suppl. Fig. S1). Continuous technical maintenance will be provided by the Institute of Medical Informatics, University of Münster. 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, Niki Tomas 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.

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

Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, et al. Ensembl 2019. Nucleic acids research 2019;47:D745-D751.

Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, et al. Ensembl 2019. Nucleic acids research 2019;47:D745-D751.

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.

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.

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.

Irie N, Weinberger L, Teng WWC, Kobayashi T, Viukov S, Manor YS, Dietmann S, Hanna JH, Surani MA. SOX17 is a critical specifier of human primordial germ cell fate. Cell 2015;160:253-268.

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.

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.

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.

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

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.

Table legend

Table 1. Information classes and items that represent publications in the MFGA.

Figure legends

Figure 1. Data sets available on the MFGA grouped by OMICS type. Currently (30 April 2020), 115 data sets corresponding to 54 publications are fully curated and publically accessible. Plot created with R (R Development Core Team, 2008).

Figure 2. 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 plots show the number of returned data sets and the frequency of gene *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 shortened and screenshot is compressed for better representability. See Suppl. Fig. S4 for full screenshot. Retrieved 30 April 2020.

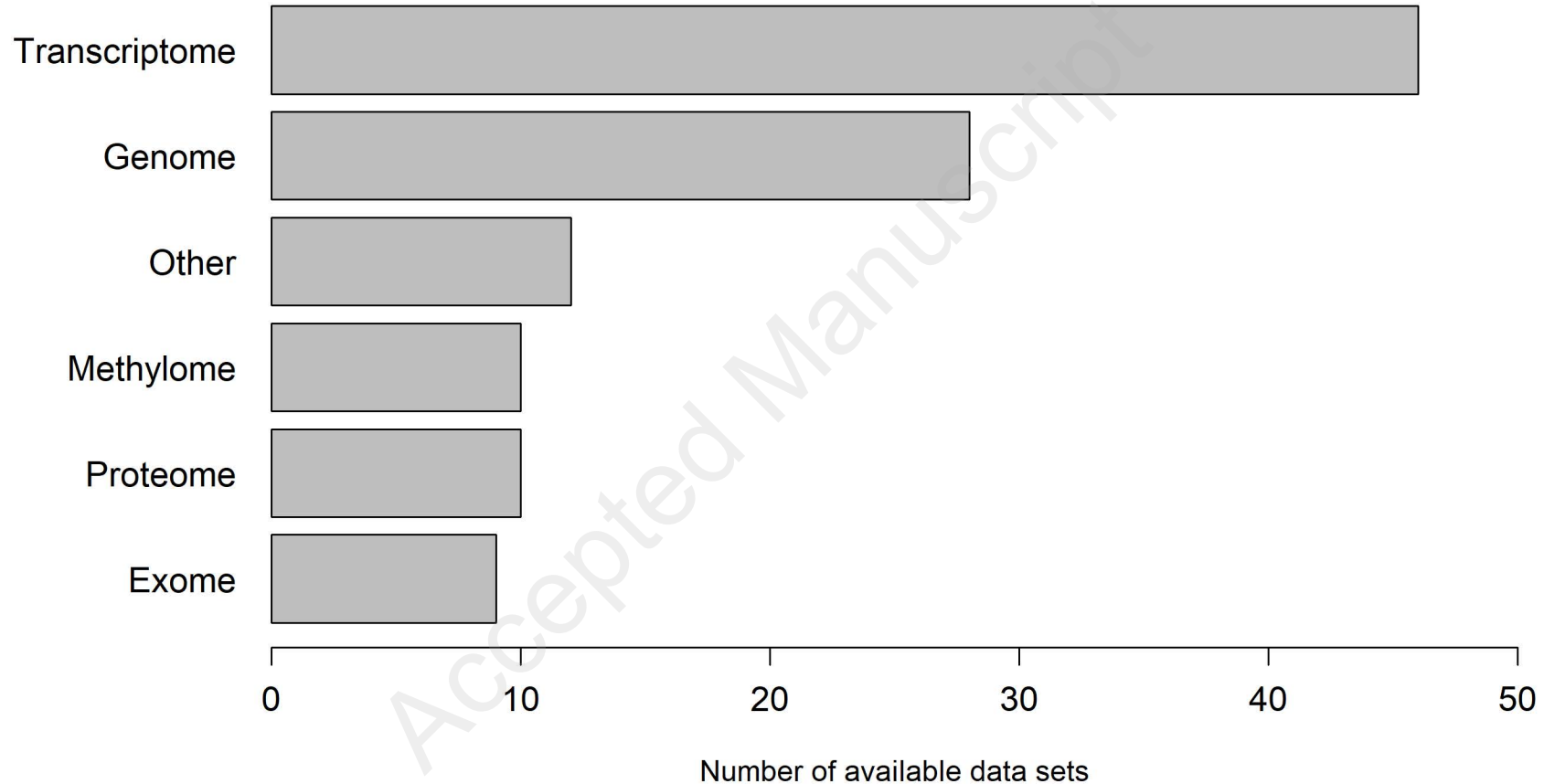
Figure 3. Example screenshot of the MFGA's functionality to represent data from publications (https://mfga.uni-muenster.de/publication/results?ssDetailsId=3968811). It shows part of supplementary table 3 of van der Bijl *et al.* (2019) in MFGA format. The table is represented in an interactive way and 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. S7. Retrieved 30 April 2020.

Figure 4. Example screenshot of the MFGA showing the overlay for additional information on genes. 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). Links for searching the gene in the MFGA database and opening the corresponding GeneCards (Stelzer *et al.*, 2016) or Ensembl id (Cunningham *et al.*, 2019) entry are provided. Additionally, gene expression in different tissues is shown based on the data from the GTEx project (Lonsdale *et al.*, 2013). Here, the gene expression plot is shown only in part. Retrieved 30 April 2020.

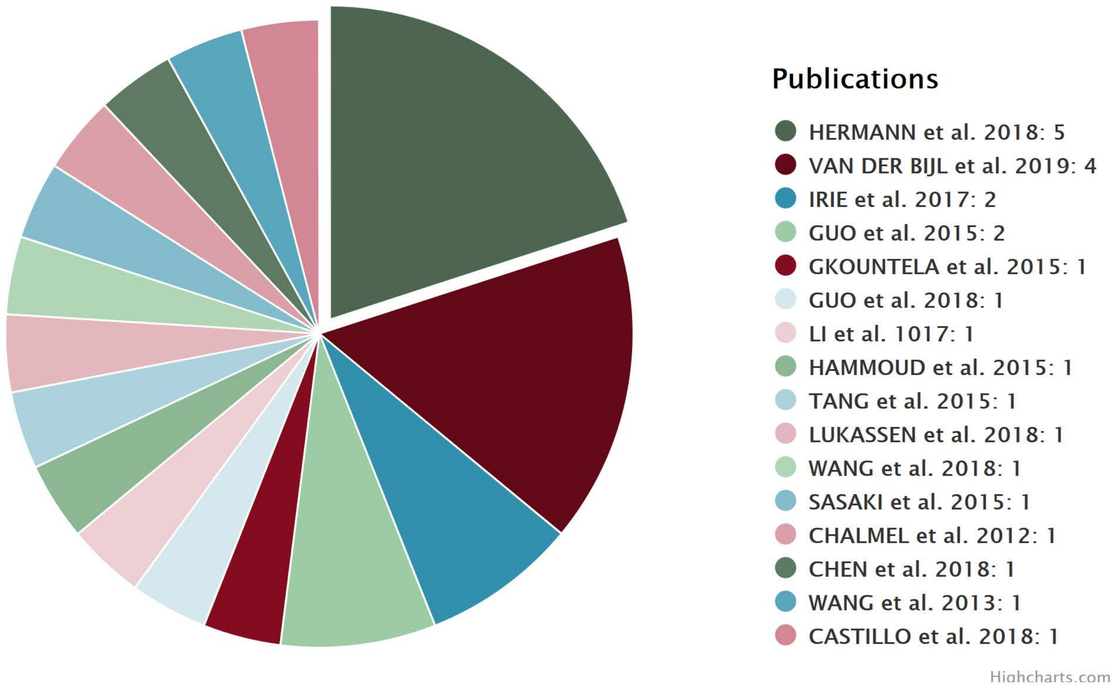
Table 1. Information classes and items that represent publications in the MFGA.

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

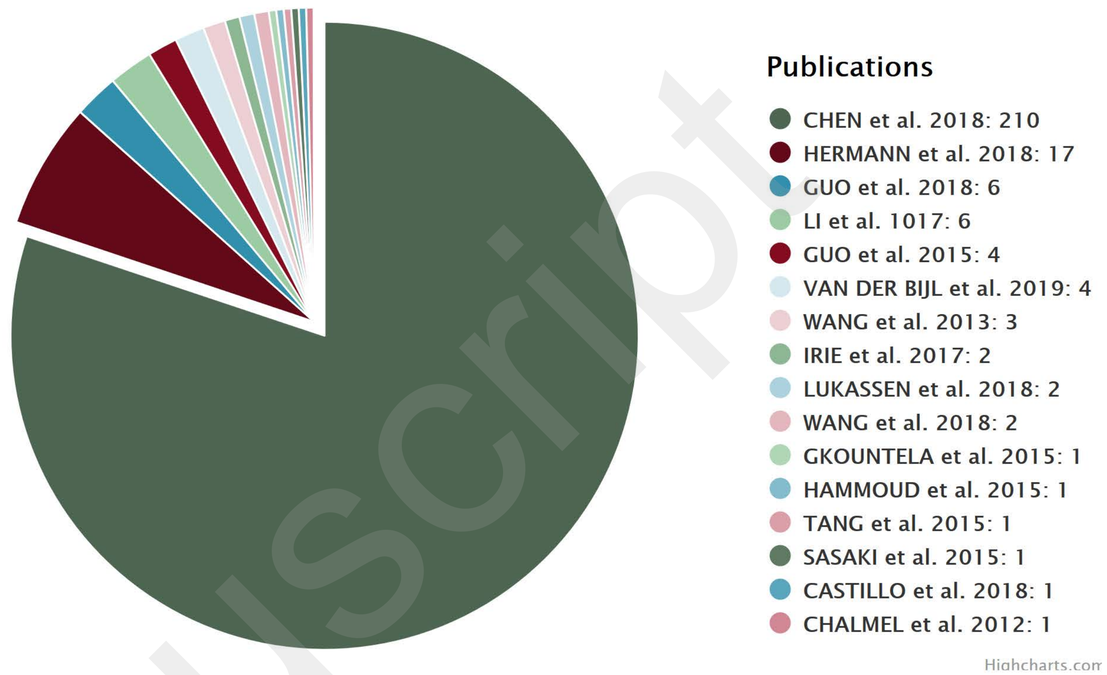
[1]*Optional items are marked with *.*



Number Of Data Sets Found Per Publication



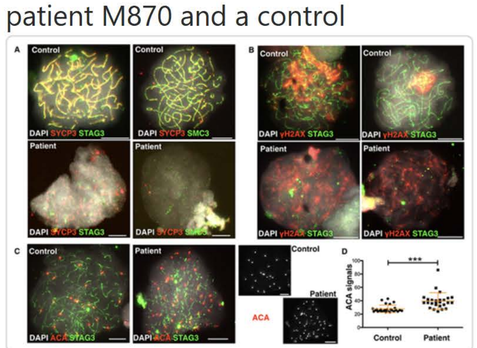
Occurrences Of Genes Per Publication



Datasets

Link	Access	Access Data In Rgv	Publication	Data Set	Or Proteins	Species	Omics	Analysis Type
			Search..	Search..	Search..	Search..	Filter	Filter
			IRIE et al. 2017, Cell.	RNA-Seq results for human PGCLCs, PGCs, and TCam-2 Seminoma cells	SOX17 BLIMP1	Human	Transcriptome	Single-cell RNA-Sequencing
			GKOUNTELA et al. 2015, Cell.	Transcriptional landmarks of human prenatal germline development		Human	Transcriptome	Bulk RNA-Sequencing

Detailed results for gene or Id search

Link	Access	Access Data In Rgv	Publication	Data Set	Result	Species	Protein	Gene	Artifact Id	Search Hits
			Search..	Search..	Search..	Filter	Search..	Search.	Search..	Search.
			WANG et al. 2018, Cell Stem Cell.	Global Transcriptional Profiling of Adult Human Testicular Cells and Cell Type Identification	Table S2.1: SPG_DEGs	Human		STAG3		1
			VAN DER BIJL et al. 2019, Human Reproduction.	Whole exome sequencing of patient M870 with STAG3 variant and meiotic arrest	Table S3: Top 10 list of the population sampling probability (PSAP) results. The two variants in the stromal antigen 3 (STAG3) gene are highlighted	Human		STAG3		2
			VAN DER BIJL et al. 2019, Human Reproduction.	Meiotic spreads and immunofluorescence staining of patient M870 with STAG3 variant and meiotic arrest	Figure 3: Analysis of chromosome features of spermatocytes from patient M870 and a control  Source VAN DER BIJL, N., et al. Mutations in the stromal antigen 3 (STAG3) gene cause male infertility due to meiotic arrest. Human Reproduction, 2019. Licensed under http://creativecommons.org/licenses/by-nc/4.0/	Human		STAG3		1

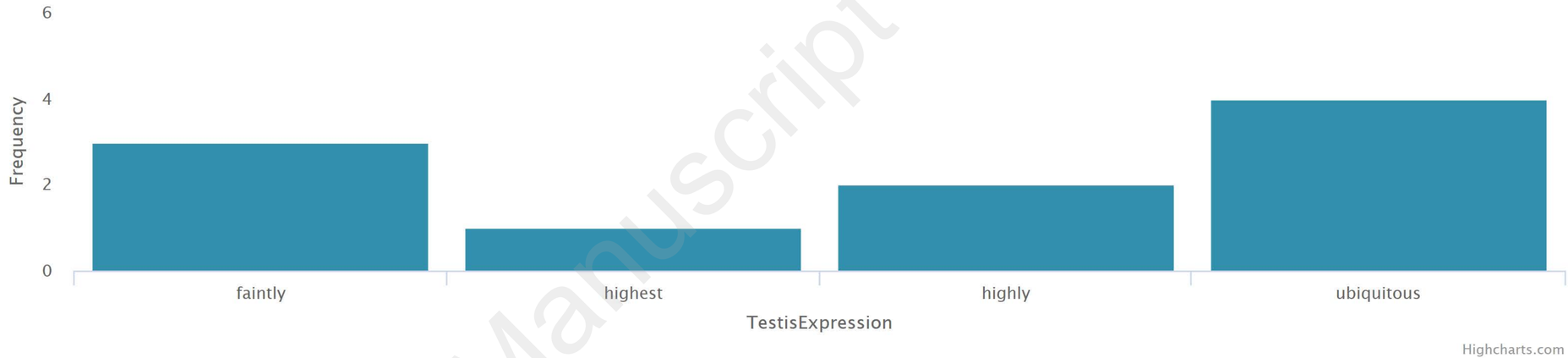
Histograms

Column: TestisExpression

Breaks: 10

Data selection: All (●) Filtered Table ( )

Hover over bars for more information



Table

Chr	Start Position	Ref	Gene	Func wg Encode Gencode Basic V19	Exonic Func wg Encode Gencode Basic V19	AAChange wg Encode Gencode Basic V19
Filter	Search..	Search.	Search.	Search..	Search..	Search..
12	121434630	-	HNF1A	exonic;intronic	frameshift insertion	HNF1A:ENST00000543427.1:exon7:c.1043_1044insTCATTCAT:p.T348fs;HNF1A:ENST0000040292
7	99796115	T	STAG3	exonic	nonsynonymous SNV	STAG3:ENST00000317296.5:exon13:c.T1262G:p.L421R;STAG3:ENST00000426455.1:exon13:c.T12
1	54605318	-	CDCP2	exonic	frameshift insertion	CDCP2:ENST00000371330.1:exon4:c.1224dupC:p.M409fs
5	175811094	-	NOP16	exonic;UTR3;downstream	frameshift insertion	NOP16:ENST00000510123.1:exon5:c.583_584insAC:p.R195fs

Location	7q22.1
Gene	STAG3 (search for STAG3 in the MFGA)
Gene full name	stromal antigen 3
Previous symbols / Alias symbols	
Gene description	stromal antigen 3 [Source:HGNC Symbol;Acc:HGNC:11356] The protein encoded by this gene is expressed in the nucleus and is a subunit of the cohesin complex which regulates the cohesion of sister chromatids during cell division. A mutation in this gene is associated with premature ovarian failure. Alternate splicing results in multiple transcript variants encoding distinct isoforms. This gene has multiple pseudogenes. [provided by RefSeq, Apr 2014]
Location type	protein-coding gene
Ensembl ID	ENSG00000066923
GeneCards	STAG3

Source: https://www.genenames.org/download/statistics-and-files/ ('Total Approved Symbols' 19.03.2020).