Title:

MetaDome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains

Authors and affiliations:

Laurens Wiel[1,2], Coos Baakman[2], Daan Gilissen[1,3], Joris A. Veltman[4,5], Gerrit Vriend[2] and Christian Gilissen[4]

1. Department of Human Genetics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, 6525 GA, the Netherlands

2. Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, 6525 GA, the Netherlands

3. Bio-informatica, HAN University of Applied Sciences, Nijmegen, 6525 EN, the Netherlands

4. Department of Human Genetics, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, 6525 GA, the Netherlands

5. Institute of Genetic Medicine, International Centre for Life, Newcastle University, Newcastle upon Tyne, NE1 3BZ, United Kingdom

Corresponding author: Christian Gilissen

Abstract (199/200)

The growing availability of human genetic variation has given rise to novel methods of measuring genetic tolerance that better interpret variants of unknown significance. We recently developed a novel concept based on protein domain homology in the human genome to improve variant interpretation. For this purpose we mapped population variation from the Exome Aggregation Consortium (ExAC) and pathogenic mutations from the Human Gene Mutation Database (HGMD) onto Pfam protein domains. The aggregation of these variation data across homologous domains into meta-domains allowed us to generate base-pair resolution of genetic intolerance profiles for human protein domains.

Here we developed MetaDome, a fast and easy-to-use web service that visualizes meta-domain information and gene-wide profiles of genetic tolerance. We updated the underlying data of MetaDome to contain information from 56,319 human transcripts, 71,419 protein domains, 12,164,292 genetic variants from gnomAD, and 34,076 pathogenic mutations from ClinVar. MetaDome allows researchers to easily investigate their variants of interest for the presence or absence of variation at corresponding positions within homologous domains. We illustrate the added value of MetaDome by an example that highlights how it may help in the interpretation of variants of unknown significance. The MetaDome web server is freely accessible at https://stuart.radboudumc.nl/metadome.

Key Words

Genetic variation; pathogenicity; web server; protein domain homology; genetic tolerance; meta-domains; gnomAD; ClinVar; Pfam

Introduction

The continuous accumulation of human genomic data has spurred the development of new methods to interpret genetic variants. There are many freely available web servers and services that facilitate the use of these data by non-bioinformaticians. For example, the ESP Exome Variant Server (Fu et al., 2012; NHLBI GO Exome Sequencing Project (ESP), 2011) and the Genome Aggregation Database (gnomAD) browser (Karczewski et al., 2017; Lek et al., 2016) help locate variants that occur frequently in the general population. These services are used for the interpretation of unknown variants based on the assumption that variants occurring frequently in the general population are unlikely to be relevant for patients with Mendelian disorders (Amr et al., 2016). There are also methods that derive information from these large human genetic databases. For example genetic intolerance, which is commonly used to interpret variants of unknown significance by assessing whether variants stand out because they occur in regions that are genetically invariable in the general population (Ge et al., 2016; Gussow et al., 2016). Examples of such methods are RVIS (Petrovski, Wang, Heinzen, Allen, & Goldstein, 2013) and subRVIS (Gussow, Petrovski, Wang, Allen, & Goldstein, 2016). The strongest evidence for the pathogenicity of a genomic variant comes from the presence of that variant in any of the clinically relevant genetic variant databases such as the Human Gene Mutation Database (HGMD) (Stenson et al., 2017) or the public archive of clinically relevant variants (ClinVar) (Landrum et al., 2016). These databases are gradually growing in the amount of validated pathogenic information.

Another way to provide evidence for the pathogenicity of a genomic variant is to observe the effect of that variant in homologous proteins across different species. Mutations at corresponding locations in homologous proteins are found to result in similar effects on protein stability

(Ashenberg, Gong, & Bloom, 2013). Finding homologous proteins is one the key applications of BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990). Transferring information between homologous proteins is one of the oldest concepts in bioinformatics, and can be achieved by performing a multiple sequence alignment (MSA) and locating equivalent positions between the protein sequences. We have previously used this concept and showed that it also holds for homologous Pfam protein domain relationships within the human genome. We found that ~71-72% of all disease-causing missense variants from HGMD and ClinVar occur in regions translating to a Pfam protein domain and observed that pathogenic missense variants at equivalent domain positions are often paired with the absence of population-based variation and *vice versa* (Wiel, Venselaar, Veltman, Vriend, & Gilissen, 2017). By aggregating variant information over homologous protein domains, the resolution of genetic tolerance per position is increased to the number of aligned positions. Similarly, the annotation of pathogenic variants found at equivalent domain positions also assists the interpretation of variants of unknown significance. We realized that this type of information could be of great benefit to the genetics community and therefore developed MetaDome.

MetaDome is a freely available web server that uses our concept of 'meta-domains' (i.e. aggregated homologous domains) to optimally use the information from population-based and pathogenic variation datasets without the need of a bioinformatics intermediate. MetaDome is easy to use and utilizes the latest population datasets by incorporating the gnomAD and ClinVar datasets.

Results

Accessibility

The MetaDome web server is freely accessible at https://stuart.radboudumc.nl/metadome. MetaDome features a user-friendly web interface and features a fully interactive tour to get familiar with all parts of the analysis and visualizations.

All source code and detailed configuration instructions are available in our GitHub repository: https://github.com/cmbi/metadome.

The underlying database: a mapping between genes and proteins

The MetaDome web server queries genomic datasets in order to annotate positions in a protein or a protein domain. Therefore, the server needs access to genomic positional information as well as protein sequence and protein domain information. The database maps GENCODE gene translations to entries in the UniProtKB/Swiss-Prot databank in a per-position manner and corresponding protein domains or genomic variation. With respect to our criteria to map gene translations to proteins (**Methods; creating the mapping database**), 42,116 of the 56,319 full-length protein-coding GENCODE Basic transcripts for 19,728 human genes are linked to 33,492 of the 42,130 Swiss-Prot human canonical or isoform sequences. Of the total 591,556 canonical and isoform sequences present in Swiss-Prot, 42,130 result from the Human species. The resulting mappings contain 32,595,355 unique genomic positions that are linked to 19,226,961 residues in Swiss-Prot protein sequences.

71,419 Pfam domains are linked to 30,406 of the Swiss-Prot sequences in our database. Of these Pfam domain instances, 5,948 are from a unique Pfam domain family and 3,334 of these families have two or more homologues and are therefore suitable for meta-domain construction. Thus, by

incorporating every protein-coding transcript, instead of only the longest ones, we increase the previously 2,750 (Wiel et al., 2017) meta-domains to 3,334. These meta-domains, on average, consist of 16 human protein domain homologues with a protein sequence length of 158 residues. **Table 1** summarizes the counting statistics for sequences, domains, etc.

Table 1. Statistics on the number of entries present in GENCODE, Swiss-Prot, and our mapping database.

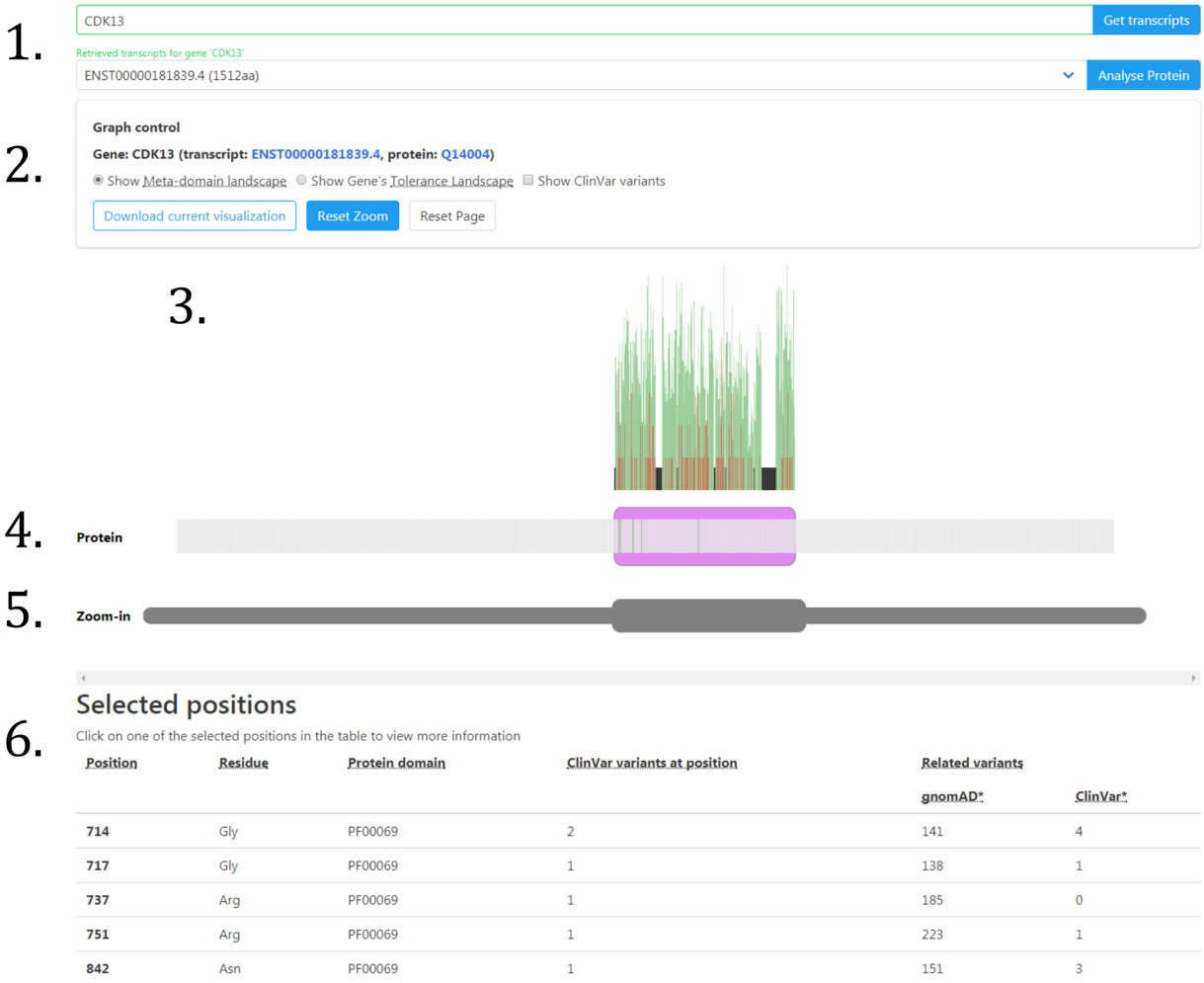
Database	What	# of entries
GENCODE	Protein-coding genes	20,345
MetaDome	Protein-coding genes	19,728
GENCODE	Protein-coding transcripts	57,005
MetaDome	Protein-coding transcripts	56,319
Swiss-Prot	Canonical and isoform protein sequences	591,556
Swiss-Prot	Human canonical and isoform protein sequences	42,130
MetaDome	Gene translations identically mapped to a canonical or isoform protein sequence	42,116
MetaDome	Canonical and isoform protein sequences	33,492
MetaDome	Pfam protein domain regions	71,419
MetaDome	Unique Pfam protein domain families	5,948
MetaDome	Unique Pfam protein domain families with two or more within-human occurrences	3,334
MetaDome	Chromosome to protein position mappings	70,261,143
MetaDome	Unique chromosome positions	32,595,355

MetaDome	Unique residues (as part of a protein)	19,226,961
MetaDome	Unique protein sequences with at least one Pfam domain annotated	30,406

How to use the MetaDome web server

At the welcome page users are offered the option to start an interactive tour or start with the analysis. The navigation bar at the top is available throughout all web pages in MetaDome and allow for further navigation to the 'About', 'Method', 'Contact' page (**Supp. Figure S1**). The user can fill in a gene symbol in the 'gene of interest' field and is aided by an auto-completion to help you find your gene of interest more easily (**Supp. Figure S2**). Clicking the 'Get transcripts' fills all GENCODE transcripts for that gene in the dropdown box. Only the transcripts that are mapped to a Swiss-Prot protein can be used in the analysis, the others are displayed in grey (**Supp. Figure S3**).

Clicking the 'Start Analysis' button starts an extensive query to the back-end of the web server for the selected transcript. Firstly, all the mappings are retrieved for the transcript of interest. Secondly, the entire transcript is annotated with ClinVar and gnomAD single nucleotide variants (SNVs) and Pfam domains. Thirdly, if there are any Pfam domains suitable for meta-domain relations then all mappings for those regions are gathered and annotated with ClinVar and gnomAD variation (**methods; Composing a meta-domain**).



Position	Residue	Protein domain	ClinVar variants at position	Related variants	
				gnomAD*	ClinVar*
714	Gly	PF00069	2	141	4
717	Gly	PF00069	1	138	1
737	Arg	PF00069	1	185	0
751	Arg	PF00069	1	223	1
842	Asn	PF00069	1	151	3

Figure 1. MetaDome web server result for the gene *CDK13*

The result provided by the MetaDome web server for the analysis of gene *CDK13* with transcript ENST00000181839.4, as provided in 1.). In 2.), there is additional information that the translation of this transcript corresponds to Swiss-Prot protein Q14004. Here also various alternative visualizations can be selected. The visualization starts by default in the 'meta-domain landscape', a mode selectable in the graph control in 2.). The landscapes are visualized in 3.), and in the meta-domain landscape the domain regions are annotated with missense variation

136 counts found in homologous domains as bar plots. The schematic protein representation, located
137 at 4.), is per-position selectable, and the domains are presented as purple blocks. Selected
138 positions are highlighted in green. The 'Zoom-in' section at 5.) features a selectable greyed-out
139 copy of schematic protein representation that can zoom-in on any part of the protein. Any
140 selected positions are in the list of selected positions in 6.). Here more information can be
141 obtained by clicking on one of these positions. A detailed description of the functionality of each
142 component is described in **Table 2.**

143

144 The web-page provided to the user as a result of the 'Analyse Protein' can best be explained
145 using an example. Therefore, we have generated this result for gene *CDK13* for transcript
146 'ENST00000181839.4' (**Figure 1**). The result page features four main components that we will
147 describe from top to bottom. Located at the top is the graph control field. Directly below the
148 graph control is the landscape view of the protein. Below the landscape view, a schematic and
149 interactive representation of the protein and an additional representation of the protein which
150 controls the zooming option. Lastly, at the bottom of the page there is the list of selected
151 positions. All of these components are interactive and the various functionalities are described in
152 **Table 2**.

153 Another way to use population-based variation in the context of the entire protein is via the
154 tolerance landscape representation in MetaDome that can be selected in the graph control
155 component (**Figure 1.2**). The tolerance landscape depicts a missense over synonymous ratio
156 (also known as K_a/K_s or d_N/d_S) over a sliding window of 21 residues over the entirety of the
157 protein of interest (e.g. calculated for ten residues left and right of each residue) based on the

158 gnomAD dataset (**methods; Computing genetic tolerance and generating a tolerance**
159 **landscape; Figure 2A**). Previously, the d_N/d_S metric has been used by others and us to measure
160 genetic tolerance and predict disease genes (Ge, Kwok, & Shieh, 2015; Gilissen et al., 2014;
161 Lelieveld et al., 2017), and it is suitable for measuring tolerance in regions within genes (Ge et
162 al., 2016).

163

164 **Table 2**. Descriptions of the various functionalities on the MetaDome result page.

Component	Functionality
Gene and transcript input field (Figure 1.1)	• Input of gene of interest • Retrieving transcripts for gene of interest • Selecting a transcript • Starting the analysis for selected transcript
Graph control field (Figure 1.2)	• Toggling between different landscape representations • Reset the zoom on the landscape • Reset the web page • Toggle ClinVar variants to be displayed in the schematic protein • Download the visual representation
Landscape view (Figure 1.3)	• Displays the meta-domain landscape • Displays the tolerance landscape
Schematic protein	• Displays a schematic representation of the gene's protein

(Figure 1.4)	with Pfam protein domains annotated • Hovering over a position displays positional information • Clicking on a position highlights the position and adds the position to the list of 'Selected Positions' • Controls the zooming of particular parts of the protein (Figure 1.5)
Selected Positions (Figure 1.6)	• Displays any positions selected in the schematic protein • Displays per selected position: if that position is part of a Pfam protein domain, any known gnomAD or ClinVar variants present at this position, and any variants that are homologously related to this position • Provides more detailed information as a pop-up when clicking on one of the positions in this list.

An example of using the MetaDome web server for variant interpretation

The MetaDome analysis result for *CDK13* (**Figure 1**) is the longest protein coding transcript for *CDK13* with a protein sequence length of 1,512 amino acids. In the resulting schematic protein representation we can observe the Pkinase Pfam protein domain (PF00069) between positions 707 and 998 as the only protein domain in this gene (**Figure 2B**). The Pkinase domain is highly prevalent throughout the human genome with as many as 779 homologous occurrences in human proteins, of which 353 are unique genomic regions. It is the 8th most occurring domain in our mapping database. The meta-domain landscape is the default view mode and shows any missense variation found in homologous domain occurrences throughout the human genome.

Population-based (gnomAD) missense variation is displayed in green and pathogenic (ClinVar) missense variation is annotated in red bars, with the height of the bars depicting the number of variants found at each position (**Figure 2B**).

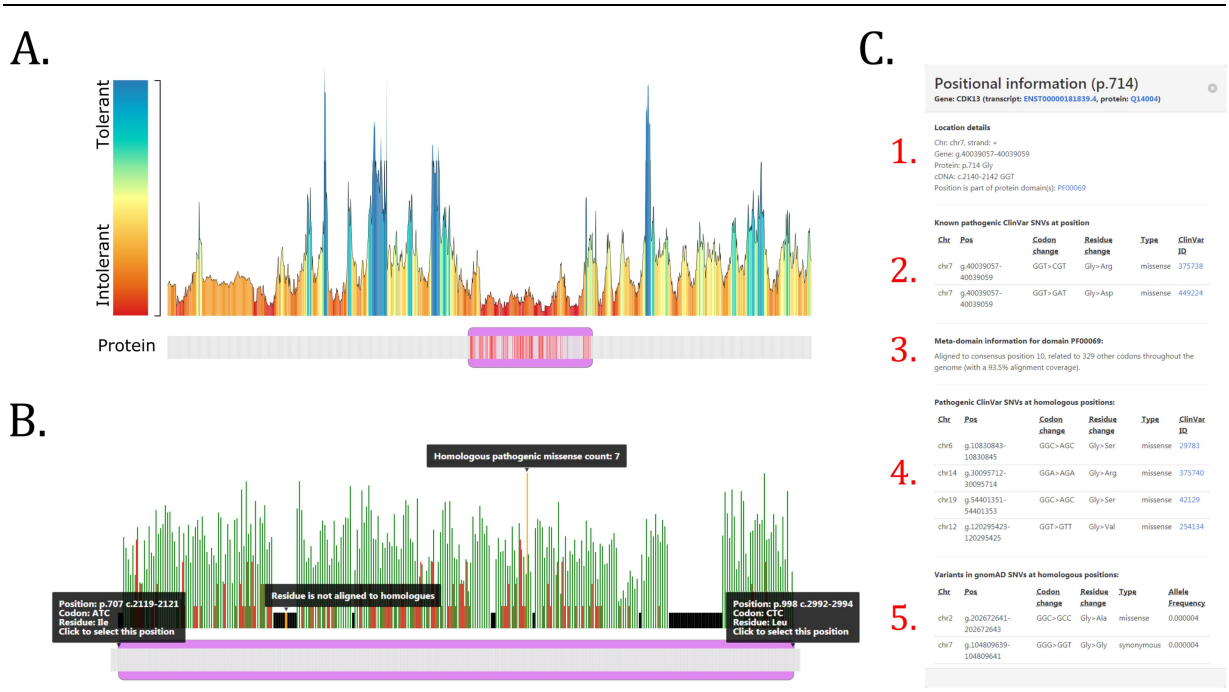


Figure 2. Examples of a MetaDome analysis for the gene *CDK13*

A.) The tolerance landscape depicts a missense over synonymous ratio calculated as a sliding window over the entirety of the protein (**methods; Computing genetic tolerance and generating a tolerance landscape**). The missense and synonymous variation are annotated from the gnomAD dataset and the landscape provides some indication of regions that are intolerant to missense variation. In this *CDK13* tolerance landscape the Pkinase Pfam protein domain (PF00069) in purple can be clearly seen as intolerant if compared to other parts in this protein. The red bars in the schematic protein representation correspond to pathogenic ClinVar variants found in this gene and in homologous protein domains. All of these variants are contained in the intolerant region of the landscape.

B.) A zoom-in on the meta-domain landscape for *CDK13*. The Pkinase Pfam protein domain (PF00069) is located between protein positions 707 and 998 and annotated as a purple box in the schematic protein representation. The meta-domain landscape displays a deep annotation of the protein domain: the green (gnomAD) and red (ClinVar) bars correspond to the amount of missense variants found at aligned homologous positions. Unaligned positions are annotated as black bars. All of this information is displayed upon hovering over these various elements.

C.) The positional information provides a detailed overview of a position from the 'Selected Positions' list, especially if that position is aligned to domain homologues. Here, for position p.Gly714 we can observe in 1.) the positional details for this specific protein position. In 2.) is any known pathogenic information for this position. We can observe here that for this position there are two known pathogenic missense variants. In 3.) meta-domain information is displayed and we can observe that p.Gly714 is aligned to consensus position 10 in the Pkinase Pfam protein domain and related to 329 other codons. This consensus position has an alignment coverage of 93.5% for the meta-domain MSA. There are also four pathogenic variants found in ClinVar on corresponding homologous positions as can be seen in 4.) and in 5.) there is an overview of all corresponding variants found in gnomAD.

If the 'Show ClinVar variants' is toggled in the graph control field, we may observe that all known pathogenic information is highlight positions in the schematic protein in a red colour and for this example is located in the protein domain (**Figure 2A**). All ClinVar variants annotated to the gene are displayed this way in red. Additionally, any ClinVar variants that are present at homologous positions are also displayed in red. In total six known disease causing SNVs are

present in the *CDK13* gene itself according to ClinVar, and these all fall within the Pkinase protein domain. All of these are missense variants. If we add variants found in homologous domains there are 64 positions with one or more reported pathogenic variants (**Supp Data S1**). Four of these positions overlap with the positions on which ClinVar variants were found in the gene itself and on position p.883 (**Supp. Figure S4**) we can observe a peak of eight missense variants annotated from other protein domains.

MetaDome helps to look in more detail to a position of interest. If we do this for protein position 714 (**Figure 2C**) in *CDK13* we find that it corresponds to consensus position 10 in the Pkinase domain (PF00069). At this position in *CDK13* there are two variants reported in ClinVar: p.Gly714Arg (ClinVar ID: 375738) submitted by (Sifrim et al., 2016), and p.Gly714Asp (ClinVar ID: 449224) submitted by GeneDX. The first is reported as a *de novo* variant and is associated to Congenital Heart Defects, Dysmorphic Facial Features, and Intellectual Developmental Disorder. For the second there is no associated phenotype provided. As MetaDome annotates variants reported at homologous positions, we can find even more information for this particular position. At the homologues aligned to this position we find a variant of identical change in *PRKD1*: p.Gly600Arg (ClinVar ID: 375740) reported as pathogenic and *de novo* in the same study (Sifrim et al., 2016). It is also associated to Congenital Heart Defects as well as associated to Ectodermal Dysplasia. There are three more reported pathogenic variants aligned to this position: *MAK*:p.Gly13Ser (ClinVar ID: 29783) associated to Retinitis Pigmentosa 62 (Özgül et al., 2011), *PRKCG*:p.Gly360Ser (ClinVar ID: 42129) associated to Spinocerebellar Ataxia Type14 (Klebe et al., 2005), and *CIT*:p.Gly106Val (ClinVar ID: 254134) associated to Microcephaly 17, primary, autosomal recessive (Özgül et al., 2011). These homologously related pathogenic variants and the severity of the associated phenotypes

234 contributes to the evidence that this particular residue may be important at this position. Further
235 evidence can be found from the fact that in human homologue domains this residue is extremely
236 conserved. There are 330 unique genomic regions encoding for a codon aligned to this position
237 (**Supp Data S2**). Only in the gene *PIK3R4* (ENST00000356763.3) does this codon encode for
238 another residue than Glycine, namely a Threonine at position p.Thr35.

239 In the same way that we explored pathogenic ClinVar variation we can also explore the variation
240 reported in gnomAD. In *CDK13* at protein position 714 there is no reported variant in gnomAD,
241 but there are homologously related variations. There are 65 missense variants with average allele
242 frequency of 1.24E-05 and 76 synonymous with average allele frequency 8.71E-03 and there is
243 no reported nonsense variation (**Supp Data S1**).

244 When we inspect the tolerance landscape for *CDK13* (**Figure 2A**) we can see that all of the
245 ClinVar variants (either annotated in *CDK13* or related via homologues) fall within the Pkinase
246 Pfam protein domain (PF00069). In addition, the protein domain can clearly be seen as more
247 intolerant to missense variation as compared to other parts of this protein, thereby supporting the
248 ClinVar variants likely pathogenic role.

249 ## Conclusion

250 The MetaDome web server combines resources and information from different fields of expertise
251 (e.g. genomics and proteomics) in order to increase the power in analysing population and
252 pathogenic variation by transposing this variation to homologous protein domains. Such a
253 transfer of information is achieved by a per-position mapping between the GENCODE and
254 Swiss-Prot databases. 79.4% of the Human Swiss-Prot protein sequences are of identical match
255 to one or more of 42,116 GENCODE transcripts. This means that 25.7% of the GENCODE

transcriptions differ in mRNA but translate to the same Swiss-Prot protein sequence. GENCODE previously reported that this is due to alternative splicing, of which a substantial proportion only affect untranslated regions (UTRs) and thus have no impact on the protein-coding part of the gene (Harrow et al., 2006).

MetaDome is especially informative if a variant of interest falls within a protein domain that has homologues. This is highly likely as 43.6% of the positions in the MetaDome mapping database are part of a homologous protein domain. Pathogenic missense variation is also highly likely to fall within a protein domain as we previously observed for 71% of HGMD and 72% of ClinVar pathogenic missense variants (Wiel et al., 2017). By aggregating variation over protein domain homologues via MetaDome, the resolution of genetic tolerance at a single base-pair is increased. Furthermore, we can obtain variation that could disrupt the functionality of a protein domain, as annotated throughout the entire human genome, which may potentially be disease-causing. It should be noted, that by aggregating genetic variation in this way the specific context such as haplotype information or interactions with other proteins may be lost. Aggregation via meta-domains only encapsulates general biological or molecular functions attributed to the domain. Nonetheless, we believe MetaDome can be used to better interpret variants of unknown significance through the use of meta-domains and tolerance landscapes as we have shown in our example.

As more genetic data accumulates in the years to come, MetaDome will become more and more accurate in predictions of intolerance at the base-pair level and the meta-domain landscapes will become even more populated with variation found in homologue protein domains. We can imagine many other ways of integrating this type of information to be helpful for variant interpretation. Future directions for the MetaDome web server could lead to machine learning

empowered variant effect prediction, or visualization of the meta-domain information in a protein 3D structure.

Methods

Software architecture of MetaDome

MetaDome is developed in Python v3.5.1 (Rossum & Drake, 2010) and makes use of the Flask framework v0.12.4 (Ronacher, 2010) for the web service part which communicates between the front-end, the back-end, and the database. The software architecture (**Supp. Fig S5**) follows the Domain-driven design paradigm (Evans, 2004). The code is open source and can be found at our GitHub repository: https://github.com/cmbi/metadome. Detailed instructions on how to deploy the MetaDome web server can be found there too.

To ensure MetaDome can be deployed to any environment and provide a high degree of modularity, we have containerized the application via Docker v17.12.1 (Hykes, 2013). We use docker-compose v1.17.1 to ensure that different containerized aspects of the MetaDome server can work together. The following aspects are containerized to this purpose: 1.) The Flask application, 2.) a PostgreSQL v10 database wherein the mapping database is stored, 3.) a Celery v4.2.0 task queue management system to facilitate the larger tasks of the MetaDome web-based user requests, 4.) a Redis v4.0.11 for task result storage, and 5.) RabbitMQ v3.7 to mediate as a task broker between client and workers. For a full overview of the docker-compose architecture we refer to **Supp. Fig. S6**.

The visualization medium of the MetaDome web server is a fully interactive and responsive HTML web page. This page is generated by the Flask framework and the navigation aesthetics

300 are made using the CSS framework Bulma v0.7.1 (Thomas, Potiekhin, Lauhakari, Shah, &
301 Berning, 2018). The visualizations of the various landscapes and the schematic protein are
302 created with JavaScript, JQuery v3.3.1, and the D3 Framework v4.13.0 (Bostock, Ogievetsky, &
303 Heer, 2011).

304 **Datasets of population and disease-causing genetic variation**

305 MetaDome makes use of single nucleotide variants (SNVs) from population and clinically
306 relevant genetic variation databases. Population variation was obtained from the gnomAD r2.0.2
307 VCF file by selecting all synonymous, nonsense, and missense variants that meet the PASS filter
308 criteria. Variants meeting the PASS criteria are considered to be true variants (Lek et al., 2016).
309 The variants in the VCF file from ClinVar release 2018 05 03 with disease-causing (Pathogenic)
310 status are used as the disease-causing SNVs in MetaDome.

311 **Creating the mapping database**

312 MetaDome stores a complete mapping between genomic, protein positions, and all domain
313 annotations (**Supp. Fig. S7**) in a PostgreSQL relational database (PostgreSQL Global
314 Development Group, 1996). This mapping is auto-generated and stored in the PostgreSQL
315 database by the MetaDome web server upon the first run. The genomic positions consist of each
316 chromosomal position in the protein-coding transcripts of the GENCODE release 19
317 GRCh37.p13 Basic set (Harrow et al., 2012). The protein positions correspond to protein
318 sequence positions in the UniProtKB/Swiss-Prot Release 2016_09 databank entries for the
319 human species (Boutet et al., 2016). These mappings are created with Protein-Protein BLAST
320 v2.2.31+ (Camacho et al., 2009) for each protein-coding translation in the GENCODE Basic set
321 to human canonical and isoform Swiss-Prot protein sequences. We exclude sequences that do not

start with a start codon (i.e. ATG encoding for methionine), or end with a stop codon. We checked if the cDNA sequence of the transcripts match the GENCODE translation via Biopython's translate function (Cock et al., 2009), if they are not identical than these are excluded too. The global information on the transcript (e.g. identifiers, sequence length) is registered in the database in the table 'genes' and, for each Swiss-Prot entry with an identical sequence match, the global information is stored in the table 'proteins'.

Next, for each identical match between translation and Swiss-Prot sequence a ClustalW2 v2.1 (Larkin et al., 2007) alignment is made between these two sequences. Each nucleotide's genomic position is mapped to the protein position and stored in the 'mappings' table. Each entry in mapping represents a single nucleotide of a codon and is linked to the corresponding entry in the 'genes' and 'proteins' table (i.e. the corresponding GENCODE translation, transcription and Swiss-Prot sequence).

Each Swiss-Prot sequence in the database is annotated via InterProScan v5.20-59.0 (Finn et al., 2017) for Pfam-A v30.0 protein domains (Finn et al., 2016) and the results are stored in the 'interpro_domains' table. After the construction of the database is finished, all meta-domain alignments can be constructed.

Composing a meta-domain

Meta-domains consist of homologous Pfam protein domain instances that are annotated using InterproScan. Meta-domains consist of at least two homologous domains. MSAs are made using a three step process. 1.) Retrieve all sequences for the domain instances, 2.) Retrieve the Pfam HMM corresponding to the Pfam identifier annotated by InterproScan, and 3.) Use HMMER 3.1b2 (Finn et al., 2015) to align the sequences from the first step. The resulting Stockholm

format MSA files can be inspected with alignment visualization software like Jalview (Waterhouse, Procter, Martin, Clamp, & Barton, 2009). In this Stockholm formatted file, all columns that correspond to the domain consensus represent the same homologous positions.

These Stockholm files are retrieved by the MetaDome web server when a user request meta-domain information for a position of their interest. Upon retrieval of this Stockholm file, the mapping database is used to obtain the corresponding genomic positions for each residue. These genomic positions are subsequently used to retrieve corresponding gnomAD or ClinVar variation.

Computing genetic tolerance and generating a tolerance landscape

The non-synonymous over synonymous ratio, or d_N/d_S score, is used to quantify genetic tolerance. This score is based on the observed (obs) missense and synonymous variation in gnomAD ($missense_{obs}\ and\ synonymous_{obs}$). This score is corrected for the sequence composition by taking into account the background (bg) of possible missense and synonymous variants based on the codon table ($missense_{bg}$ and $synonymous_{bg}$):

$$d_N/d_S = \frac{missense_{obs}/missense_{bg}}{synonymous_{obs}/synonymous_{bg}}.$$

The tolerance landscape computes this ratio as a sliding window of size 21 (i.e. ten residues before and ten after the residue of interest) over the entirety of the gene's protein. The edges (e.g. start and end) are therefore a bit noisy as they are not the result of averaging over a full length window.

Acknowledgements

This work was in part financially supported by grants from the Netherlands Organization for Scientific Research (916-14-043 to C.G. and 918-15-667 to J.A.V.), and from the Radboud Institute for Molecular Life Sciences, Radboud university medical center (R0002793 to G.V.). We thank Hanka Venselaar for her critically reading of the manuscript.

URLs

MetaDome web service: https://stuart.radboudumc.nl/metadome

GENCODE: https://www.gencodegenes.org/

InterPro: https://www.ebi.ac.uk/interpro/

Pfam: https://pfam.xfam.org/

gnomAD: http://gnomad.broadinstitute.org/

ClinVar: https://www.ncbi.nlm.nih.gov/clinvar/

GitHub repository: https://github.com/cmbi/metadome

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amr, S. S., Al Turki, S. H., Lebo, M., Sarmady, M., Rehm, H. L., & Abou Tayoun, A. N. (2016). Using large sequencing data sets to refine intragenic disease regions and prioritize clinical variant interpretation. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, (July), 1–9. https://doi.org/10.1038/gim.2016.134

Ashenberg, O., Gong, L. I., & Bloom, J. D. (2013). Mutational effects on stability are largely conserved during protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(52), 21071–6. https://doi.org/10.1073/pnas.1314781111

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D^3: Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2301–9. https://doi.org/10.1109/TVCG.2011.185

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A. J., … Xenarios, I. (2016). UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Methods in Molecular Biology (Clifton, N.J.)*, *1374*, 23–54. https://doi.org/10.1007/978-1-4939-3167-5_2

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

397 Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … de Hoon, M. J.
398 L. (2009). Biopython: freely available Python tools for computational molecular biology
399 and bioinformatics. *Bioinformatics (Oxford, England)*, *25*(11), 1422–3.
400 https://doi.org/10.1093/bioinformatics/btp163

401 Evans, E. (2004). *Domain-driven design: tackling complexity in the heart of software*. Addison-
402 Wesley Professional.

403 Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., … Mitchell, A.
404 L. (2017). InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids*
405 *Research*, *45*(D1), D190–D199. https://doi.org/10.1093/nar/gkw1107

406 Finn, R. D., Clements, J., Arndt, W., Miller, B. L., Wheeler, T. J., Schreiber, F., … Eddy, S. R.
407 (2015). HMMER web server: 2015 update. *Nucleic Acids Research*, *43*(W1), W30–W38.
408 https://doi.org/10.1093/nar/gkv397

409 Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., … Bateman, A.
410 (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic*
411 *Acids Research*, *44*(D1), D279–D285. https://doi.org/10.1093/nar/gkv1344

412 Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., … Akey, J. M. (2012).
413 Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants.
414 *Nature*, *493*(7431), 216–220. https://doi.org/10.1038/nature11690

415 Ge, X., Gong, H., Dumas, K., Litwin, J., Phillips, J. J., Waisfisz, Q., … Shieh, J. T. C. (2016).
416 Missense-depleted regions in population exomes implicate ras superfamily nucleotide-
417 binding protein alteration in patients with brain malformation. *Npj Genomic Medicine*, *1*(1),

16036. https://doi.org/10.1038/npjgenmed.2016.36

Ge, X., Kwok, P.-Y., & Shieh, J. T. C. (2015). Prioritizing genes for X-linked diseases using population exome data. *Human Molecular Genetics*, *24*(3), 599–608. https://doi.org/10.1093/hmg/ddu473

Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., van de Vorst, M., van Bon, B. W. M., Willemsen, M. H., … Veltman, J. A. (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature*, *511*(7509), 344–347. https://doi.org/10.1038/nature13394

Gussow, A. B., Petrovski, S., Wang, Q., Allen, A. S., & Goldstein, D. B. (2016). The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biology*, *17*(1), 9. https://doi.org/10.1186/s13059-016-0869-4

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.-K., Chrast, J., … Guigo, R. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biology*, *7 Suppl 1*(Suppl 1), S4.1-9. https://doi.org/10.1186/gb-2006-7-s1-s4

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., … Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, *22*(9), 1760–1774. https://doi.org/10.1101/gr.135350.111

Hykes, S. (2013). Docker. San Francisco: Docker, Inc. Retrieved from https://www.docker.com/

Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., … MacArthur, D. G. (2017). The ExAC browser: Displaying reference data information

439 from over 60 000 exomes. *Nucleic Acids Research*, *45*(D1), D840–D845.
440 https://doi.org/10.1093/nar/gkw971

441 Klebe, S., Durr, A., Rentschler, A., Hahn-Barma, V., Abele, M., Bouslam, N., … Stevanin, G.
442 (2005). New mutations in protein kinase Cγ associated with spinocerebellar ataxia type 14.
443 *Annals of Neurology*, *58*(5), 720–729. https://doi.org/10.1002/ana.20628

444 Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., … Maglott, D. R.
445 (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic*
446 *Acids Research*, *44*(D1), D862-8. https://doi.org/10.1093/nar/gkv1222

447 Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H.,
448 … Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*(21),
449 2947–2948. https://doi.org/10.1093/bioinformatics/btm404

450 Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., …
451 MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans.
452 *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

453 Lelieveld, S. H., Wiel, L., Venselaar, H., Pfundt, R., Vriend, G., Veltman, J. A., … al., et.
454 (2017). Spatial Clustering of de Novo Missense Mutations Identifies Candidate
455 Neurodevelopmental Disorder-Associated Genes. *The American Journal of Human*
456 *Genetics*, *8*(0), 52–56. https://doi.org/10.1016/j.ajhg.2017.08.004

457 Li, W. H., Wu, C. I., & Luo, C. C. (1985). A new method for estimating synonymous and
458 nonsynonymous rates of nucleotide substitution considering the relative likelihood of
459 nucleotide and codon changes. *Molecular Biology and Evolution*, *2*(2), 150–74.

https://doi.org/10.1128/AAC.03728-14

NHLBI GO Exome Sequencing Project (ESP). (2011). Exome Variant Server. Retrieved May 14, 2015, from http://evs.gs.washington.edu/EVS/

Özgül, R. K., Siemiatkowska, A. M., Yücel, D., Myers, C. A., Collin, R. W. J., Zonneveld, M. N., … Corbo, J. C. (2011). Exome sequencing and cis-regulatory mapping identify mutations in MAK, a gene encoding a regulator of ciliary length, as a cause of retinitis pigmentosa. *American Journal of Human Genetics*, *89*(2), 253–264. https://doi.org/10.1016/j.ajhg.2011.07.005

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., & Goldstein, D. B. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genetics*, *9*(8), e1003709. https://doi.org/10.1371/journal.pgen.1003709

PostgreSQL Global Development Group. (1996). PostgreSQL. PostgreSQL Global Development Group. Retrieved from https://www.postgresql.org/

Ronacher, A. (2010). Flask. Retrieved from http://flask.pocoo.org/

Rossum, G. Van, & Drake, F. L. (2010). Python Tutorial. *History*, *42*(4), 1–122. https://doi.org/10.1111/j.1094-348X.2008.00203_7.x

Sifrim, A., Hitz, M.-P., Wilsdon, A., Breckpot, J., Turki, S. H. Al, Thienpont, B., … Hurles, M. E. (2016). Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nature Genetics*, *48*(9), 1060–5. https://doi.org/10.1038/ng.3627

Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., … Cooper, D. N.

(2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics*, *136*(6), 1–13. https://doi.org/10.1007/s00439-017-1779-6

Thomas, J., Potiekhin, O., Lauhakari, M., Shah, A., & Berning, D. (2018). *Creating Interfaces with Bulma*. (T. Mott & D. Berning, Eds.). Santa Rosa: Bleeding Edge Press. Retrieved from https://bleedingedgepress.com/

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189–1191. https://doi.org/10.1093/bioinformatics/btp033

Wiel, L., Venselaar, H., Veltman, J. A., Vriend, G., & Gilissen, C. (2017). Aggregation of population-based genetic variation over protein domain homologues and its potential use in genetic diagnostics. *Human Mutation*, (May), 1–10. https://doi.org/10.1002/humu.23313

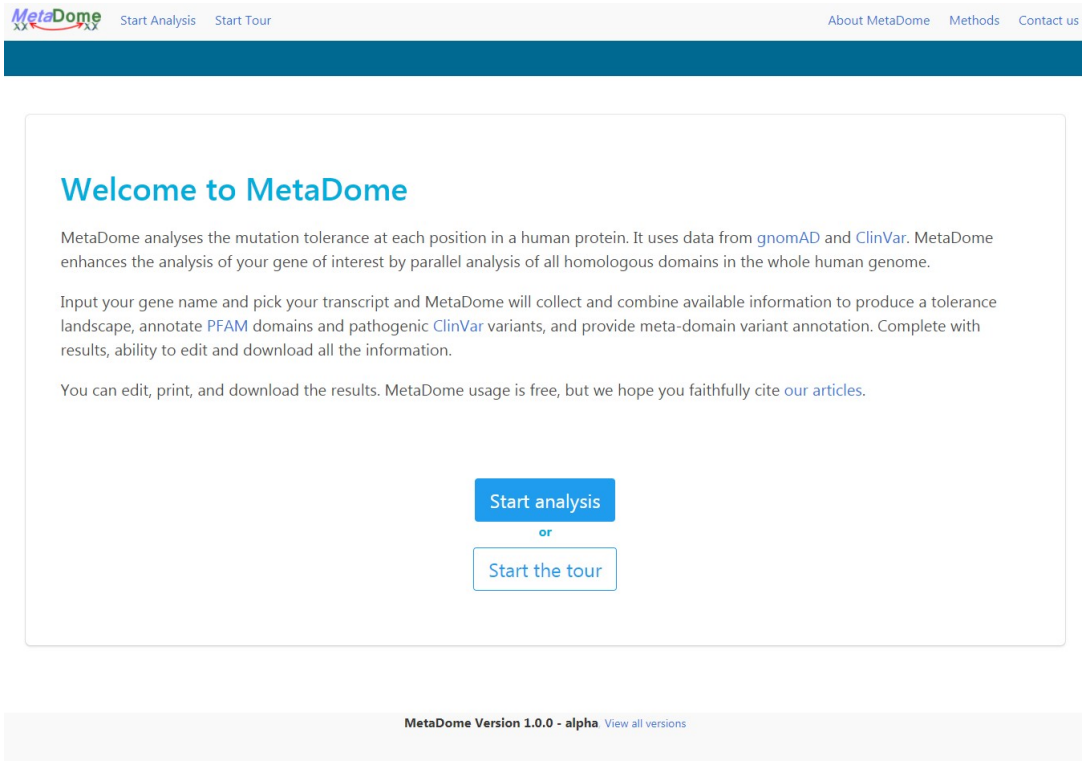
Yang, & Bielawski. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, *15*(12), 496–503. https://doi.org/10.1016/S0169-5347(00)01994-7

Yang, Z., Swanson, W. J., & Vacquier, V. D. (2000). Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Molecular Biology and Evolution*, *17*(10), 1446–55. https://doi.org/10.1093/oxfordjournals.molbev.a026245

Supplemental Figures

503

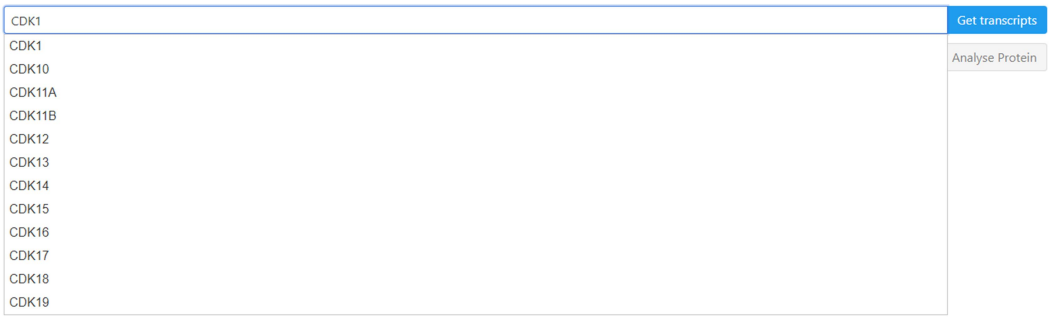
504 **Supp Fig. S1. Welcome page**



505

506 The welcome page of MetaDome is the entry point to the rest of the web server. Here the
507 navigation bar is located at the top that eases navigation throughout the rest of the web pages.
508 From here the user can start the interactive tour or an analysis.

509 **Supp Fig. S2. Gene symbol auto-completion**

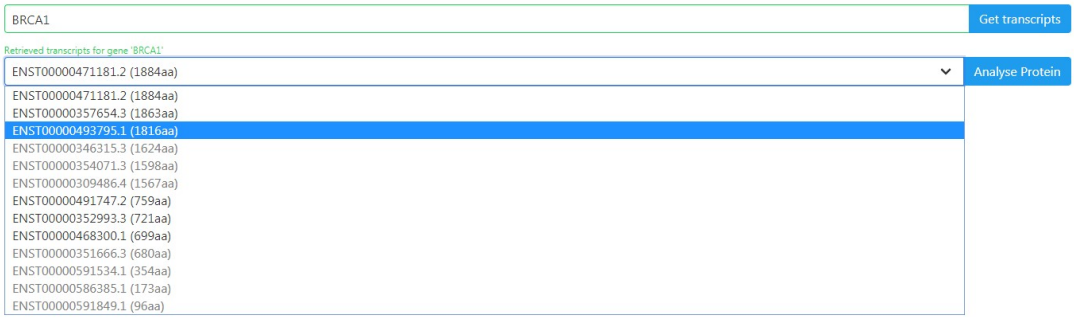


510

511 The auto-completion of gene symbols is based on all gene symbols present in the mapping
512 database. The auto-completion will start once the user has entered three characters in the input
513 field and is interactive (e.g. you can click on the gene that is of your interest.

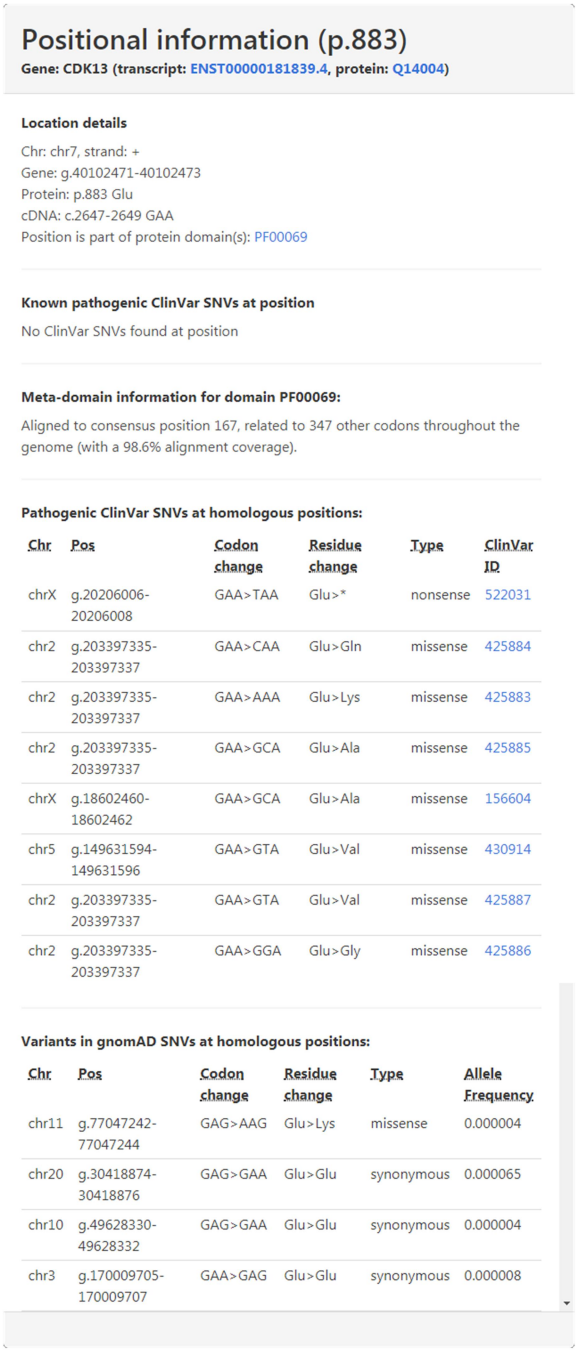
514

Supp Fig. S3. Dropdown box containing transcript identifiers



The result of clicking the 'Get transcripts' button for the *BRCA1* gene. Here a success message is displayed in green, which means that the gene is present in the database. The resulting transcripts are ordered by amino acid (aa) length and the longest transcripts are at the top. The greyed-out transcripts are not suitable for further analysis in the web server.

Supp Fig. S4. Dropdown box containing transcript identifiers



Positional information (p.883)

Gene: CDK13 (transcript: ENST00000181839.4, protein: Q14004)

Location details

Chr: chr7, strand: +
Gene: g.40102471-40102473
Protein: p.883 Glu
cDNA: c.2647-2649 GAA
Position is part of protein domain(s): PF00069

Known pathogenic ClinVar SNVs at position

No ClinVar SNVs found at position

Meta-domain information for domain PF00069:

Aligned to consensus position 167, related to 347 other codons throughout the genome (with a 98.6% alignment coverage).

Pathogenic ClinVar SNVs at homologous positions:

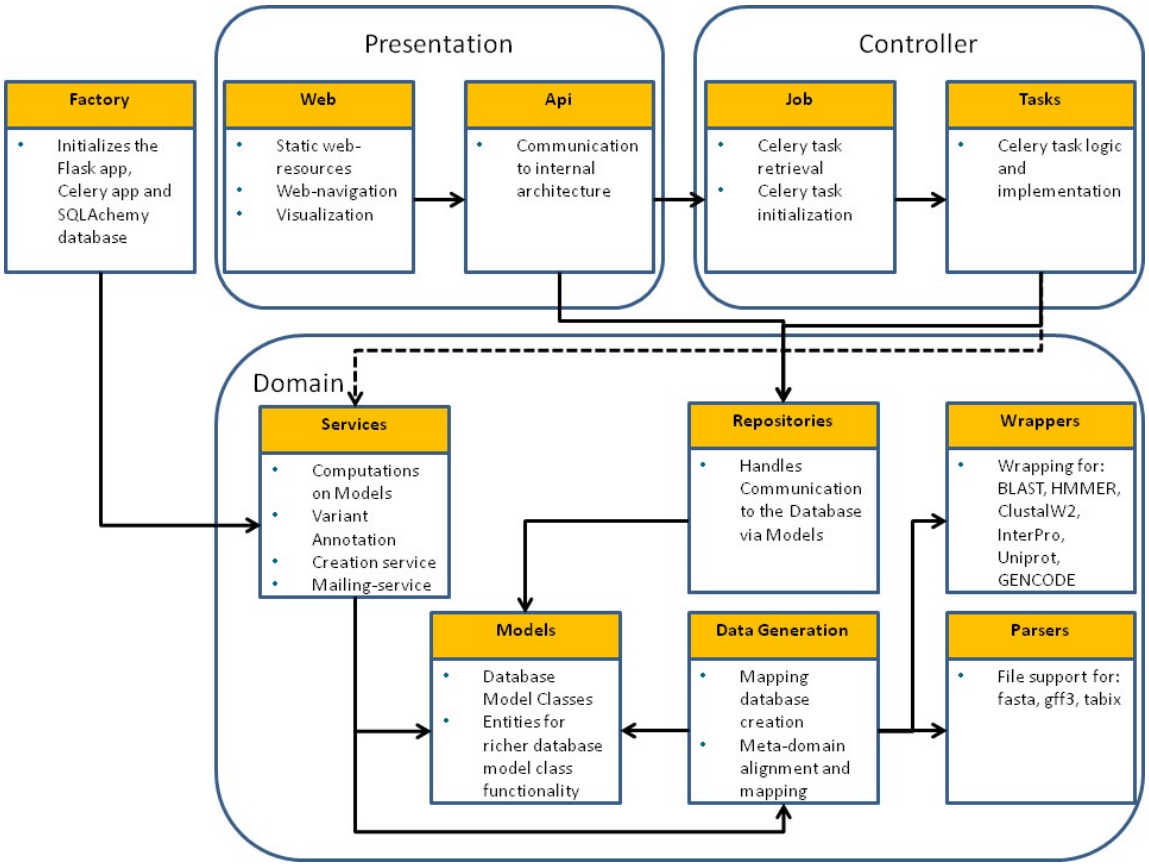
Chr	Pos	Codon change	Residue change	Type	ClinVar ID
chrX	g.20206006-20206008	GAA>TAA	Glu>*	nonsense	522031
chr2	g.203397335-203397337	GAA>CAA	Glu>Gln	missense	425884
chr2	g.203397335-203397337	GAA>AAA	Glu>Lys	missense	425883
chr2	g.203397335-203397337	GAA>GCA	Glu>Ala	missense	425885
chrX	g.18602460-18602462	GAA>GCA	Glu>Ala	missense	156604
chr5	g.149631594-149631596	GAA>GTA	Glu>Val	missense	430914
chr2	g.203397335-203397337	GAA>GTA	Glu>Val	missense	425887
chr2	g.203397335-203397337	GAA>GGA	Glu>Gly	missense	425886

Variants in gnomAD SNVs at homologous positions:

Chr	Pos	Codon change	Residue change	Type	Allele Frequency
chr11	g.77047242-77047244	GAG>AAG	Glu>Lys	missense	0.000004
chr20	g.30418874-30418876	GAG>GAA	Glu>Glu	synonymous	0.000065
chr10	g.49628330-49628332	GAG>GAA	Glu>Glu	synonymous	0.000004
chr3	g.170009705-170009707	GAA>GAG	Glu>Glu	synonymous	0.000008

The positional information overview provides detailed positional information for position p.Glu883. We can observe that it is aligned to consensus position 167 in the Pkinase Pfam protein domain. This consensus position has 98.6% alignment coverage. On this position no ClinVar pathogenic variants are found in *CDK13*, but on homologous positions there are eight pathogenic ClinVar variants annotated.

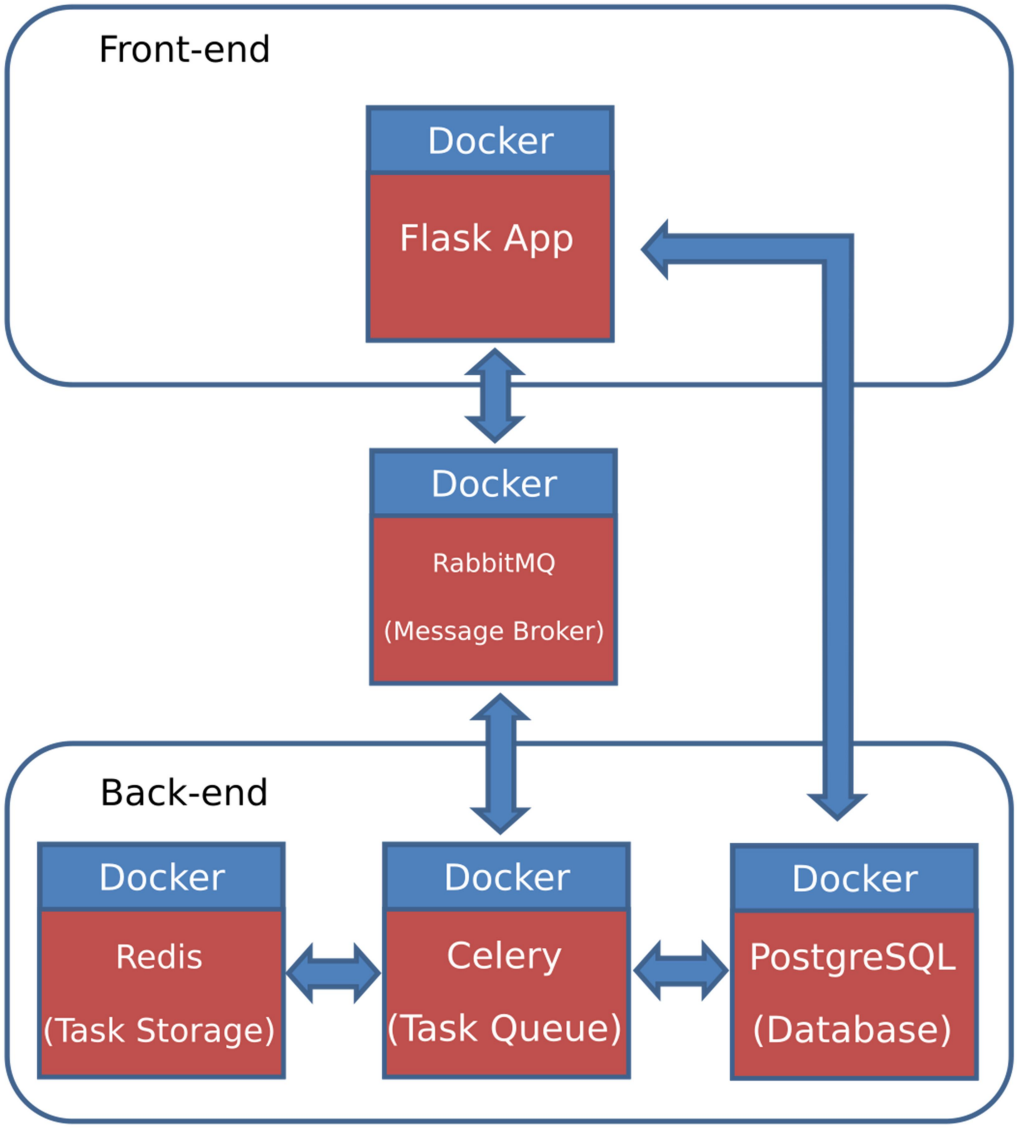
Supp Fig. S5. The software architecture of the MetaDome web server



A high-level overview of the software architecture of the MetaDome web server. The architecture follows closely the Domain-driven design paradigm. In this architecture we make use of three distinct layers with separated responsibilities. The Presentation layer is responsible for the user interface and external communication handling. This layer takes care of all the visualizations in the web interface and any user input that needs further validation or that requires starting a celery job. In the controller layer all communication is handled from the MetaDome server to the Celery task queue. Job here instantiates and checks the status of jobs in the Celery queue that are initialized from Tasks. The Domain layer contains all the domain-knowledge of the MetaDome application. Here Repositories is the outside connection for approaching the internal database, so no domain knowledge is needed by the Presentation layer.

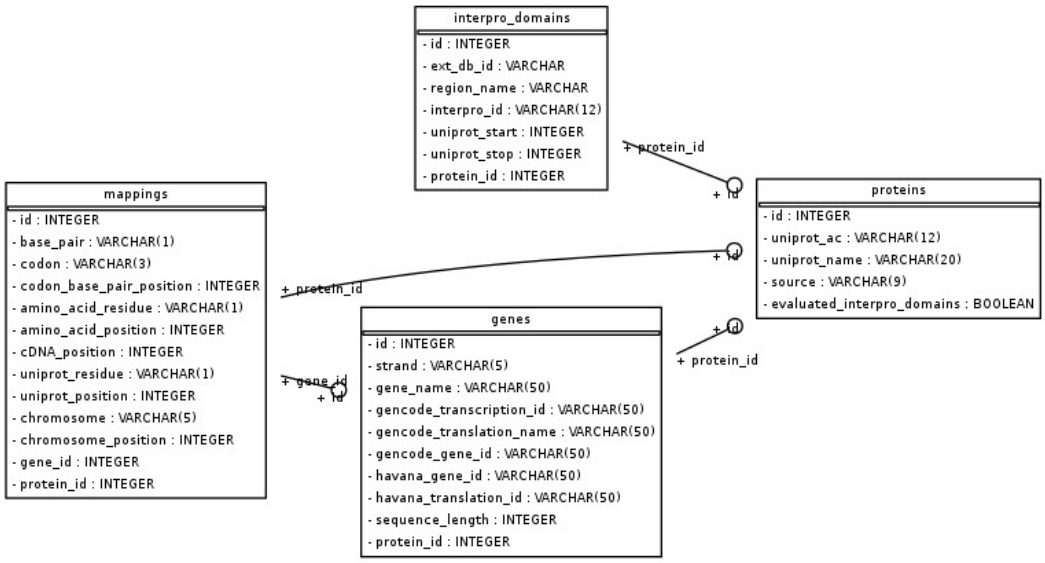
The services take care of various computations on the data from Models and contains functionality to create the database via Data Generation. The Factory is outside the layers, but initializes the configuration to run the MetaDome server (e.g. connections to the PostgreSQL and Celery Docker containers and start of the Flask web service).

Supp Fig. S6. The internal docker network architecture of the MetaDome web server



Metadome makes use of several docker containers that work jointly via docker-compose to provide various degrees of functionality. The Flask App docker container is responsible for hosting the MetaDome visual interface and at the first run it will create the mapping database. The Flask docker container communicates via the RabbitMQ service, which serves as a message broker, to the Celery docker container. This Celery container keeps track of the large tasks and the results of those tasks are temporarily stored in the Redis server. The PostgreSQL container houses the database containing the mapping between GENCODE transcripts and Swiss-Prot.

554 **Supp Fig. S7. The internal database architecture for the MetaDome web server**



555

556 The backbone of the MetaDome web server is a relational PostgreSQL database. This schematic
557 representation of the tables present an overview of the relations in the data and what type of
558 information is stored for each data entry. The table ‘mappings’ contain entries of a per-
559 chromosome position (per gene) entries with codon to amino acid residue information. This is
560 done for each of the nucleotides in a codon. The ‘genes’ table represent information from each of
561 the transcripts present in the GENCODE database and the ‘proteins’ table correspond to global
562 information on entries from the UniProtKB/Swiss-Prot databank. The ‘interpro_domains’ table
563 can support any type of interpro domain, but for the purpose of MetaDome contain solely the
564 Pfam protein domains. This database brings together genomic and protein positions that can be
565 further used for annotation purposes.

566