Collaboratively charting the gene-to-phenotype network of human congenital heart defects

Barriot, R.[1,2,*], Breckpot, J.[3,*], Thienpont, B.[3,*], Van Vooren, S.[1], Coessens, B.[1], Tranchevent, L.-C.[1], Van Loo, P.[1], Gewillig, M.[4], Devriendt, K.[3,+], Moreau, Y.[1,+,#]

[1] ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

[2] Laboratoire de Microbiologie et Génétique Moléculaires, 118 Rte de Narbonne, 31000 Toulouse, France

[3] Center for Human Genetics, University of Leuven, Herestraat 49, 3000 Leuven, Belgium

[4] Department of Pediatric Cardiology, University Hospital Leuven, Herestraat 49, 3000 Leuven, Belgium

* These authors contributed equally to this work

[+] Shared senior authors

[#] To whom correspondence should be addressed

Anonymous reviewer logins

Usernames: Reviewer1, Reviewer2, Reviewer3

Password: CHDWiki

Abstract

How to efficiently integrate the daily practice of molecular biologists, geneticists, and clinicians with the emerging computational strategies from systems biology is still much of an open question. Using a collaborative Wiki, we developed a knowledge base and gene prioritization portal aimed at mapping genes involved in congenital heart defects (CHDs) and untangling their relations with corresponding

human phenotypes. This portal is not only an evolving community repository of current knowledge on the genetic basis of congenital heart defects, but also a collaborative environment for the study of candidate genes potentially implicated in CHDs – in particular by integrating recent strategies for the statistical prioritization of candidate genes. It thus serves the broad community that is facing CHDs, ranging from the pediatric cardiologist and clinical geneticist to the basic investigator of cardiogenesis. We demonstrate the usefulness of this approach in the prioritization of a set of chromosomal regions identified by microarray comparative genomic hybridization (array CGH) in patients suffering from CHDs. We readily identify genes with proven involvement in CHDs, as well as some genes most likely involved in CHD, such as *HAND2* and *BMP4*. Of broad interest to the biological community, we argue that specialized, collaborative knowledge and analysis portals will play a significant role in systems biology studies of numerous complex biological processes.

CHDWiki is accessible at http://www.esat.kuleuven.be/~bioiuser/chdwiki

Introduction

Recently, Wiki technology – inspired by the well-known Wikipedia encyclopedia – has been proposed as a potential strategy for the collaborative development of biological knowledge bases [1-7]. Although a 'Wikipedia for Genes' is likely to emerge, a number of challenges remain. First, classical Wiki technology in itself (based on free text) is unsuitable for developing genetic knowledge bases because of the imperative need for structured information. Hence, Wiki platforms for genetic knowledge bases need to provide a strong framework for integration with classical database technology. Wikiproteins already implements this need at a high level by abstractly linking concepts, such as proteins and biological processes. Second, and probably foremost, each community uses specific

terminology, has specific goals, and uses specific data and tools. Such specificity cannot be addressed in a generic Wikipedia for Gene and requires tailored solutions implementing different levels of specialization. Third, Wiki technology does not in itself support downstream analysis of the information gathered in the Wiki.

Going beyond knowledge gathering, integrative data analysis strategies have been proposed recently for the prioritization of genes potentially involved in a given biological process, phenotype, or disease[8-10]. Nevertheless, there is clearly a gap between such advanced (and somewhat complex) analysis strategies and actual wet lab work. A similar gap can be observed between those strategies and clinical genetics where increasingly complex molecular data needs to be interpreted towards the diagnosis of constitutional disorders. To bridge this gap and bring integrative analysis strategies into practice, we integrate a candidate gene prioritization method[10] and browsing of networks of gene interactions[11-13] into the Wiki platform.

We therefore propose to combine Wiki technology, databases of genomic and phenomic information, and data analysis tools into a Wiki portal that supports the need of a specialized community. In particular, we describe a Wiki portal for the genetic study of congenital heart defects (CHD), termed *CHDWiki*. CHDs are the major cause of mortality in newborns in the developed world, but despite this manifest importance, the majority of CHDs still has an unknown etiology. In some instances, specific genetic and environmental factors have been shown to cause CHDs. A review of the etiology of CHDs is available through CHDWiki (http://www.esat.kuleuven.be/~bioiuser/chdwiki/index.php/CHD:Review).

The CHDWiki portal focuses on mapping out the gene network leading to human CHD phenotypes. It supports both genetic and molecular biology research that

aims at hunting for CHD genes, as well as clinical research that aims at identifying and interpreting genetic aberrations in patients suffering from well-characterized CHDs.

Methods

Knowledge acquisition

To build a set of most currently known gene-phenotype links, OMIM and MEDLINE were manually searched by an expert in the field, for genes that are linked with any of 139 relevant cardiac defect phenotypes listed among the internationally used CHD codes from the Association for European Paediatric Cardiology (AEPC). The use of this specialized ontology maximizes the relevance of the collected information to the CHD community and improves the consistency of this information. Relevant genes and mutations were selected and their corresponding cardiac phenotype were manually gathered and described based on the available literature. The level of support for a gene-phenotype link was defined by its incidence and the number of independent publications reporting it. We only considered such links confirmed if at least two reports from independent groups described the incidence of CHD in patients with a mutation to be greater than 1%. Moreover, the support for the link between every single gene mutation and CHD type was further characterized based on the genetic evidence (inheritance and incidence), *in silico* predictions, and the functional studies (*in vitro* analysis and animal models) described in the study.

To build a set of most currently known chromosomal regions linked to CHDs, MEDLINE was searched for imbalances detected by molecular karyotyping, breakpoints of balanced chromosome aberrations or linkage regions that are linked with CHDs. This data complements at a much higher resolution the CHD

regions identified by Van Karnebeek et al. (1999)[14], which were based on reported cytogenetically visible chromosomal aberrations.

CHDWiki is conceived to allow straightforward inclusion of published and unpublished data from all collaborators. All clinical (cardiac and non-cardiac) and molecular data are rendered anonymous and collected in a standardized manner, either directly in the CHDWiki database (for genes, translocation breakpoints and linkage regions) or (for well-delineated chromosomal imbalances) using another tool designed for this purpose, CHDBench (http://tomcat.esat.kuleuven.be/chdbench). Consent for submitting unpublished data from the patient or his legal representative is explicitly required, and was assumed to be obtained for published data that is included. Ethical approval for the incorporation of patient data was obtained from the Ethics Committee K.U.Leuven (S51093).

Platform development

CHDWiki is based on the MediaWiki engine initially developed for the Wikipedia project. We implemented a generic extension that allows registering specific components for the management of structured data and for the on-the-fly execution of analysis tools.

The benefits of databases are manifold and become apparent when providing different views on the same data. For instance, it allows providing the detailed list of genes linked to CHDs, as well as, the list of CHDs having linked genes. Databases also solve consistency issues, for example when a link is added or updated between a gene and a specific CHD, then both the gene page and the CHD page instantly reflect this change. The principles of the generic extensions are the following. Specific components can subscribe to pages so that the Wiki

engine executes these components when the page is rendered, or they can explicitly be called from within the page through the use of a specific tag.

For example, to include the list of CHDs linked to *JAG1*, the generic extension first calls the `chdsForGenes` component with parameter *JAG1* to retrieve the data in the form of (variable, value) pairs. Second, the extension retrieves the `chdsForGeneTemplate` layout template which is stored as Wiki text in a standard Wiki page. Third, it replaces the variables by their actual values. Eventually, the resulting Wiki text is rendered by the MediaWiki engine.

The simplicity of the generic extension mechanism makes it both flexible and powerful. Specific components already available include (1) numerous data retrievers from our local databases, (2) a chromosome map summarizing genes and genomic regions linked to CHDs, (3) pie charts generation, (4) gene network visualization and exploration, and (5) candidate gene prioritization. This variety of components illustrates the versatility of the approach. For instance, pie charts are easily included by calling the lightweight `pieChart` component with the list of slices (name and value pairs), while the prioritization component consists in a complete web application specifically tuned towards prioritizing CHD genes. More generally, the generic extension proved successful in the fast development of new components working as wrappers for databases, web services, command line tools, and DAS servers[15]. Also, to speed up the development time for structured data update and interaction, the extension implements a generic mechanism to easily specify web forms in Wiki text which are pre-filled and handled by registered components.

In addition, CHDWiki interacts with a patient data repository, CHDBench (http://tomcat.esat.kuleuven.be/chdbench), for managing patient data published in the literature, and a DAS server[16] feeding CHD genes and genomic regions has

been set up allowing one to loop from CHDWiki to the Ensembl genome browser and back.

When more Wiki portals such as CHDWiki are available, the problems of interoperability of these systems and integration of stored knowledge can be managed through standard protocols such as DAS for data access, web services for the programmatic use of knowledge, or dedicated Application Programming Interfaces (API) that will have to be further specified by the community. The component oriented architecture of CHDWiki will make such future developments easy to implement.

Prioritization of candidate genes for CHD loci identified by array CGH

Fifteen well-delineated genome regions that were found to be imbalanced in CHD patients were selected for further analysis: 10 from a previously published study[17] and 5 imbalances we detected during a follow-up study. Seven regions that contain a known CHD gene served as positive controls (Table 1). Of important note, these positive controls also correspond to the cases that can be convincingly resolved from a clinical point of view, see Discussion. Of the remaining 8 regions, we excluded 2 patients carrying multiple chromosomal imbalances from further analysis, while the 6 other chromosomal imbalances were used to identify novel putative CHD genes. This strategy has previously been successful, for example, in studies of the recurrent 22q11 deletion that led to the identification of *TBX1* mutations causing CHDs and DiGeorge syndrome[18]. The same example also demonstrates the challenges associated to this strategy, since many other candidate genes within the deletion interval had to be investigated in model organisms before identifying TBX1 – thus requiring much time and resources for the identification[19]. Therefore, to enhance candidate gene selection in our regions,

we applied the gene prioritization strategy (Endeavour) which is incorporated in CHDWiki. The results from this prioritization are provided as supplementary data.

The 6 known CHD genes and 25 high-ranking genes selected from the 6 regions containing no known CHD gene were further analyzed in zebrafish (Table 1 and Table 2). Their expression in the developing zebrafish heart was evaluated at 3 to up to 6 different developmental stages that are key for cardiac development: 12 to 18 hours post fertilization (hpf): heart cell specification and initiation of migration to the midline, 22hpf: at the midline, heart cone stage, 30 to 36 hpf: heart elongation and looping, initiation of chamber specification and valve formation, 48 hpf: valve formation. The expression was analyzed by a standard method of colorimetric whole-mount *in situ* hybridizations using DIG-labeled antisense RNA probes. RNA Probes were produced by *in vitro* transcription from Sp6 or T7-tagged DNA that was PCR amplified from a cDNA pool. This pool was reverse transcribed from an RNA mix that was extracted using TRIzol (Invitrogen, Paisley, UK) from zebrafish embryos at the 6 different defined developmental stages (12, 18, 22, 30, 36 and 48 hpf). Nested PCR was applied when PCR using a single primer pair was unable to produce a single product. When no product was obtained after nested PCR, we assumed that the transcript was not present in the cDNA pool and that the gene was therefore not expressed at the relevant stages. All experiments were carried out using the wild-type AB zebrafish line. TBX1 expression in the developing zebrafish has often been described[20,21], so that we did not reanalyze it in this study.

Overview of CHD data

The results of the knowledge acquisition are described in Table 3 and Table 4, and are visualized on an interactive chromosomal map in CHDWiki (Figure 1). They represent a unique repository of human genetic data for CHDs that describes both

the heart phenotype as well as the genetic lesion with a granularity of detail that was unavailable so far. For each gene or phenotype, CHDWiki provides a pie diagram that graphically represents the spectrum of related non-syndromic cardiac phenotypes or mutated genes, followed by a detailed overview of the studies defining the mutational spectrum of these gene-phenotype links. It allows for the addition of a free text description of any aspects of the gene that the contributor consider as relevant. Additionally, this portal offers a graphical overview of the protein interaction partners and external links to both human and non-human genome browsers and model organism databases.

Prioritization and assessment of *HAND2* and *BMP4* as CHD candidates

First, all genes from selected regions containing known CHD genes were prioritized. In all 6 regions, the gene prioritization tool readily ranked at least one known CHD gene first among all other imbalanced genes (Table 1). Analyzing their expression in the developing zebrafish demonstrated that three of them, *NKX2-5, TBX1,* and *NOTCH1* have a specific expression in the developing zebrafish heart (i.e., a restricted expression pattern that includes the developing heart). *NKX2-5* is a well-known cardiac marker and labels the cardiac cells throughout zebrafish development. *NOTCH1* is expressed at late stages of cardiac development in the developing atrioventricular valve. Since *NKX2-5* and *NOTCH1* mutations cause isolated CHDs (i.e., non-syndromic), the remainder of the features present in the patients with the imbalance must be caused by other genes in the region, and these thus represent contiguous gene syndromes. A similar observation is described in the del7q11.2 causing Williams syndrome, where haploinsufficiency of the elastin gene causes the CHD, while other genes are related to the other phenotypic manifestations[22]. *EHMT1*, *ATRX*, *CREBBP*, *NSD1,* and *FBN2* are genes that cause syndromic CHDs. These appeared to be expressed throughout the

developing zebrafish without a specific expression restricted to the developing zebrafish heart. This pattern is consistent with the multisystem involvement of the syndromes caused by mutations in these genes. Nevertheless, genes that cause syndromic CHDs when mutated (such as *TBX1*, which causes DiGeorge syndrome) can have specific expression in the developing heart.

In a second step, genes affected by the six selected imbalances not containing known CHD genes were prioritized in an attempt to find novel CHD genes (described in supplementary data: homes.esat.kuleuven.be/~bioiuser/chdwiki/index.php/Private:Prioritization). In 4 out of the 6 regions, no gene showed specific expression in the developing heart, while some showed ubiquitous expression. This suggests that the unknown gene is part of the syndromic subset: they are causing CHDs as well as the other abnormalities that cause the syndromic phenotype observed in the patients. In the 2 remaining regions, a highest-ranking gene was found to be specifically expressed in the developing heart. *HAND2* in 4q34.1 is expressed (Figure 2) at 12 hours post-fertilization (hpf) in the lateral plate mesoderm where the cardiac cells become specified, and later in development (22-48hpf) it is expressed in the primitive heart and the pharyngeal arches that give rise to the outflow tract and large vessels in humans. *BMP4* in 14q22 is specifically expressed at 18hpf in the specified cardiac cells (Figure 3) as they migrate towards the midline, later in the developing heart and at 36hpf at the regions where valve development initiates.

Discussion

Wiki portals

We argue that, next to the development of a Wikipedia for Gene, another major application of Wiki-like collaborative technology is the development of Wiki

knowledge and analysis portals. In many research areas, the body of common knowledge is usually scattered across reviews, original articles, and genomic databases (MEDLINE, Ensembl and UCSC genome browsers, Entrez, OMIM, and many others). A structured Wiki allows the integration of all these data into a view that is centered on the needs of a community. A large number of bioinformatics analysis methods are available, yet a significant gap remains towards their integration within the daily practice of most biologists. By tightly integrating a knowledge base with dedicated analysis tools, a Wiki portal provides a natural stepping stone for biologists to use advanced data analysis techniques. While the development of data analysis portals has been a long-standing aim of bioinformatics, few convincing applications with a clear impact on biology have been demonstrated since the development of homology-based methods[23,24]. We believe that the causes of this lack of progress are mainly (1) lack of fine-tuning to the needs of a specific biological research area and (2) complexity of data analysis strategies that appear overwhelming to biologists. We believe that Wiki portals alleviate those problems in several ways: (1) lowering the threshold towards data analysis by a tight integration with a knowledge base that is immediately useful to any biologist from the community and (2) development of data analysis strategies sufficiently user friendly to be within reach of the majority of biologists (such as prioritization strategies, network browsing, and so on).

Impact on knowledge acquisition and exchange

In the presented CHDWiki, we compiled all information available in the literature and published data on CHD genetics to construct a collaborative portal, introducing a new type of review of the literature, which is dynamic and evolving. It readily provides researchers, geneticists, physicists, and cardiologists a direct access to the most comprehensive knowledge on CHD genetics. An immediate impact of this methodology is how it could complement and enhance the process

of reviewing the literature. A first clear advantage of Wikis is their dynamic nature. They can accompany a review article so that information can be kept up-to-date after the publication of the original review. A second advantage is that they can link to actual data (i.e., appropriate entries in the genome browsers and cross-links to other databases and relevant data sets). A third advantage is obviously their collaborative nature where the expertise of numerous researchers can be pooled together to obtain extensive peer-reviewed knowledge.

The impact of disease-oriented Wikis is quite significant for both clinicians and researchers. For clinicians, it allows finding out instantly if a gene or region has been linked to the phenotype of their patients. For instance, the finding that a chromosomal imbalance in a CHD patient affects an annotated CHD gene or region provides conclusive evidence for its causality, thus facilitating diagnosis and therapy. In the specific case of our study of array CGH patients, this corresponds to the 6 positive controls for which a known CHD gene can be readily identified in the region and thus for which the genetic cause of the clinical phenotype can be convincingly established. For the researcher, several novel regions recurrently linked to CHDs in patients emerge from the compiled genomic data, such as 1p36.3, 6q21, 15q26, 17q21.3, and 22q12 (Figure 1). These provide an entry point to find novel genes linked to human CHDs. Indeed, in the CHDWiki, advanced analysis capabilities were exploited for gene hunting: genes located in indels found in CHD patients (Table 1 and Table 2) were prioritized using the integrated prioritization tool. Genes linked to CHDs and affected by an indel were effectively prioritized. For regions not containing known CHD genes, high ranking genes were further analyzed for their expression in developing zebrafish embryos. Amongst those genes, *BMP4* and *HAND2* showed a specific expression in the developing zebrafish heart (Figure 2 and Figure 3), suggesting that reduced dosage of these genes causes the CHD observed in the patients. Additionally, both genes have

been shown to be required for cardiac development in a dosage-dependent manner in mice[25,26]. While further studies are underway to obtain conclusive evidence for their involvement in CHDs, their potential implication in CHD is readily annotated in CHDWiki and is available to the community.

Specialized Wiki portals like CHDWiki may play a key role towards a universal 'Wikipedia of Genes'. Content from specialized Wikis can be compiled into a Gene Wikipedia. There might be some challenges with such an integration, so that an alternative model would be a federation of Wikis, where no centralized Gene Wikipedia exists, but where a hub/search engine passes queries to registered Wikis and the results are aggregated. The latter solution seems more likely since it is more efficient to develop a specialized Wiki portal such as CHDWiki providing specific tools for the community and focusing on a particular area. For instance, the Wikipedia page on Tetralogy of Fallot is directed towards a large audience with sections such as the symptoms, diagnosis and treatments, while the CHDWiki page shortly describes this phenotype and provides extensive knowledge on the genetics of this CHD (genes and mutations, patient reports). Another example is from Wikiproteins which aims at being exhaustive about available protein knowledge, where the GATA4 entry currently links to several species and provides general information (function, localization, structure) but where its link to CHDs (ASD, Tetralogy of Fallot, …) is briefly and incompletely mentioned. Use of standards and ontologies for query and exchange (possibly in the line of Semantic Web technology[27] will be paramount to ensuring portals interoperability and minimizing annotation ambiguity (i.e., semantic integrity). For instance, the use of AEPC codes for ASD distinguishes Atrial Septum Defects from Autism Spectrum Disorders.

Wiki portals also perfectly fit with the philosophy of Distributed Annotation Systems. Not only can the data of any gene – or genomic feature – be distributed among portals and collaboratively annotated, but the genome browser can also serve as an entry point to the Wikis. CHD researchers who enable the CHDWiki tracks in a genome browser directly visualize any information compiled in CHDWiki.

Collaborative aspects

An important asset of specialized Wiki portals is that they could be a solution to the long-term maintenance of biological knowledge bases. Such long-term maintenance is a major recurring issue for the biological and bioinformatics community because of the persistent lack of long-term funding. We envision a model for community-driven knowledge bases where the original portal development is funded through a classical grant, but where the curation of the knowledge base is gradually shifted from the core development team to the community at large over the period of the grant (3 to 5 years). Data curation should be included into the original development cycle until a critical mass of information is reached that guarantees adoption by the community, as has been done for CHDWiki (described in the knowledge acquisition section). After curation has been shifted to the community, technical portal maintenance can be minimal.

Wikis provide also an effective solution to the enduring problem of unpublishable or negative results (for example of mutation analyses). These can be highly valuable for other research teams pursuing similar paths of investigation, or contain relevant information below statistical significance. Wiki portals like CHDWiki are a natural repository for such findings. Also classical collaborative features, such as a mailing list (e.g., for collaboration requests, job postings, event announcements and exchange of biological material such as DNA, cell lines, or

cardiac tissue) and a directory of researchers are basic but valuable tools such a portal can offer.

Outstanding issues

A number of technical and social issues may require significant further technology developments. The first one is access control. Given the functionalities provided by a Wiki portal, it comes close to a collaborative laboratory notebook. However, research teams will understandably be reluctant to share their experimental data prior to publication. Access control restrictions (i.e., deciding to provide access to part of the information to only a selected group of people) will require some further development of the Wiki platforms, although this is basic a technical problem that is routinely solved in database systems.

A second issue is quality control. Wikipedia is based on a model of peer-editing, which is generally effective but can sometimes lead to conflicts between contributors (edit wars). Such conflicts are likely to arise when conflicting scientific hypotheses and interpretations collide. Additionally, conflicts are also likely regarding scientific attribution and priority. For smaller Wikis, a ‘benevolent dictator’ who oversees any conflict can solve those problems. For larger Wikis, a hierarchy of editorial roles (contributor, reviewer, and technical, associate, and executive editors) might become necessary.

A third issue is credit assignment. Scientific credit is an essential driver and one cannot expect researchers to contribute a significant proportion of their time if no credit is attributed to them. For smaller Wikis, citation of a key publication describing the Wiki will be sufficient (exactly like for specialized databases) because key contributors will be among the authors. For larger Wikis, a direct crediting system might be difficult to establish. Significant contributions to Wikis (as contributor or editor) should be recognized as relevant scientific work and

therefore appear in the track record of a researcher. Related to the issue of credit is the issue of citation. As some Wiki entries may become key scholarly references, proper citation to accurate dates or versions might be needed in scholarly work because entries change over time.

Regarding software development, each knowledge portal is likely to have highly specific aspects (type of information, source databases, ontologies or analysis tools) that make the development of a single off-the-shelf solution (similar to MediaWiki and other Wiki software) highly unlikely. A more realistic solution will lie in the development of generic tools that enable the flexible construction of such Wikis by embedding generic modules for ontology management, XML data representation, visualization, database query, and data analysis. The goal of such software will be to maximally speed up the development of Wikis while minimizing functional constraints. The current version of our Wiki framework has been made available as an open source project (http://www.mediawiki.org/wiki/Extension:Inout).

The future we envision is one where a specialized community “swarms” around a Wiki portal that provides most of the knowledge, data, and analysis tools needed to support its experimental work. Via this portal, the community can collaboratively and incrementally chart complex networks involved in biological processes, phenotypes, and diseases. Collaboration and efficient access to knowledge, data, and tools will significantly speed up experimental research.

Acknowledgements

This research was supported by the Research Council KUL (GOA AMBioRICS, CoE EF/05/007 SymBioSys, PROMETA, several PhD/postdoc & fellow grants, GOA 2006/12), FWO [PhD/postdoc grants, projects G.0241.04 (Functional Genomics),

G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM)], G.0254.05 (Genetics of human heart develoment), IWT (PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame, SBO-MoKa, TBM Endometriosis), the Belgian Federal Science Policy Office [IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011), IUAP P5/25 (Molecular Pathology of Genetic Diseases) and the EU-RTD (ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain, FP6-STREP Strokemap). P.V.L. is supported by a postdoctoral research fellowship, J.B. by a PhD fellowship and K.D. is a senior clinical investigator of the Research Foundation-Flanders (FWO).

References

Tables

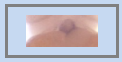
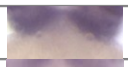
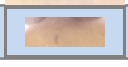
imbalance (number of genes)	reference (PMID)	CHD genes (prioritized rank)	Human phenotype (MIM number)	zebrafish ortholog ue	Heart expression (36-48hpf)
del5q23 (125)	16954676	*FBN2* (1)	congenital contractural arachnodactyly (121050)	*fbn2*	
del5q35.1 (45)	16954676	*NKX2.5* (1)	ASD with AV conduction defects (108900)	*nkx2-5*	
del5q35.3 (129)	16954676	*NSD1* (1)	Sotos syndrome (117550)	*nsd1*	
del9q34.3 (106) dup20q13.33	16954676	*EHMT1* (5)	chromosome 9q subtelomeric deletion syndrome (610253)	*ehmt1a*	
				ehmt1b	
		NOTCH1 (1)	aortic valve disease (109730)	*notch1a*	
				notch1b	
dup16p13.3 (3)	17702016	*CREBBP* (1)	Rubinstein-Taybi syndrome (180849)	*crebbp*	
dup22q11.2 (135)	16954676	*TBX1* (1)	microduplication 22q11.2 (608363)	*tbx1*	heart & branchial arches
dupXq21.1 (5)	17579672	*ATRX* (1)	α-thalassemia/mental retardation syndrome, X-linked (301040)	*atrx*	

Table 1: patients carrying imbalances encompassing known CHD genes and expression these CHD genes in the developing zebrafish heart

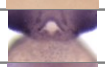
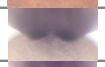
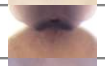
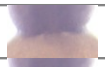
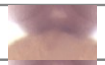
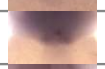
Imbalance (number of genes)	reference (PMID)	genes	zebrafish ortholog ue	Heart Expression (36-48 hpf)
del1p36.33	17337261	*DVL1*	*dvl1*	
		HES4	*hes4*	
		SKI	*skia*	
			skib	
del4q34	16954676	*HAND2*	*hand2*	
		HMGB2	*hmgb2a*	
			hmgb2b	
		VEGFC	*vegfc*	
del14q22q23.1	17384091	*ARID4A*	*arid4a*	
		BMP4	*bmp4*	
		DLG7	*dlg7*	
		OTX2	*otx2*	
dup17q25.3	NA	*No ID*	*NA*	
		AATK	*aatka*	
			aatkb	
		ACTG1	*NA*	
		CSNK1D	*csnk1d*	
		FOXK2	*foxk2a*	
			foxk2b	
		MAFG	*mafg*	
del22q12.2	17384091	*THOC5*	*thoc5*	
		EWSR1	*ewsr1a*	
			ewsr1b	
		KREMEN1	*kremen1*	
delXp22	17256798	*RAI2*	*rai2*	
		REPS2	*reps2*	

Table 2: patients carrying imbalances not encompassing known CHD genes and expression of high-ranked genes in the developing zebrafish heart

entity	entries
genes	54
CHDs	69
linkage regions	8
balanced chromosomal aberrations	17
Van Karnebeek and Hennekam 1999 regions	46
indel patients	85
indel regions	93
references	175

Table 3: number of entities (genes, CHDs, ...) present in CHDWiki

	CHD
genes	136
linkage regions	9
balanced chromosomal aberrations	19
Van Karnebeek and Hennekam 1999 regions	85
Patients	149
studies (mutation screens)	154
mutations	234

Table 4: number of links between CHDs and currently managed entities (genes, regions, ...)

Figures

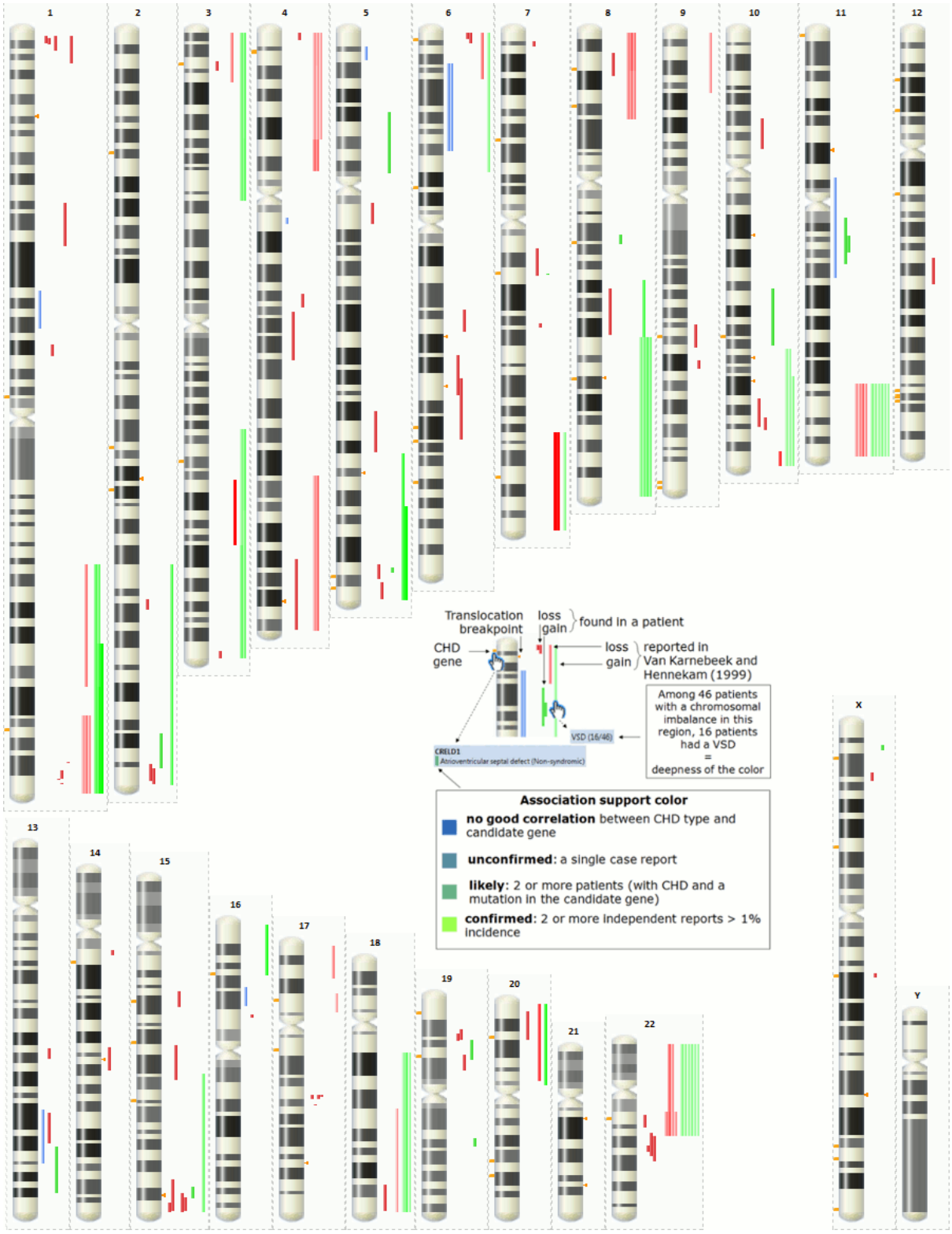


Figure 1: overview of CHD data plotted on a human karyogram. A dynamic and interactive version is available at http://www.esat.kuleuven.be/~bioiuser/chdwiki/index.php/CHD:Map

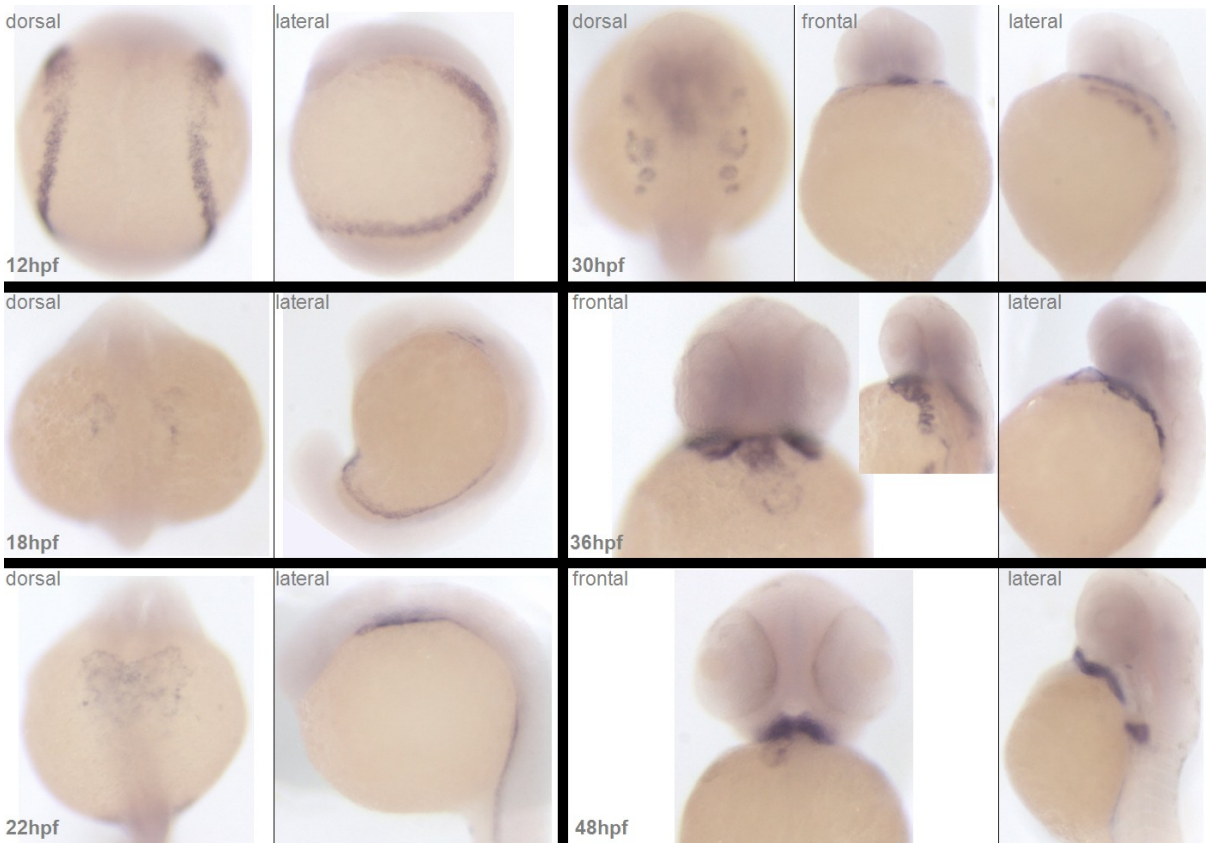


Figure 2: mRNA expression of *hand2* in the developing zebrafish at different developmental stages. At 12hpf, *hand2* is expressed in the lateral plate mesoderm that houses the cardiac progenitor cells. Later in development, *hand2*-expressing cells are found in the developing heart and pharyngeal apparatus.

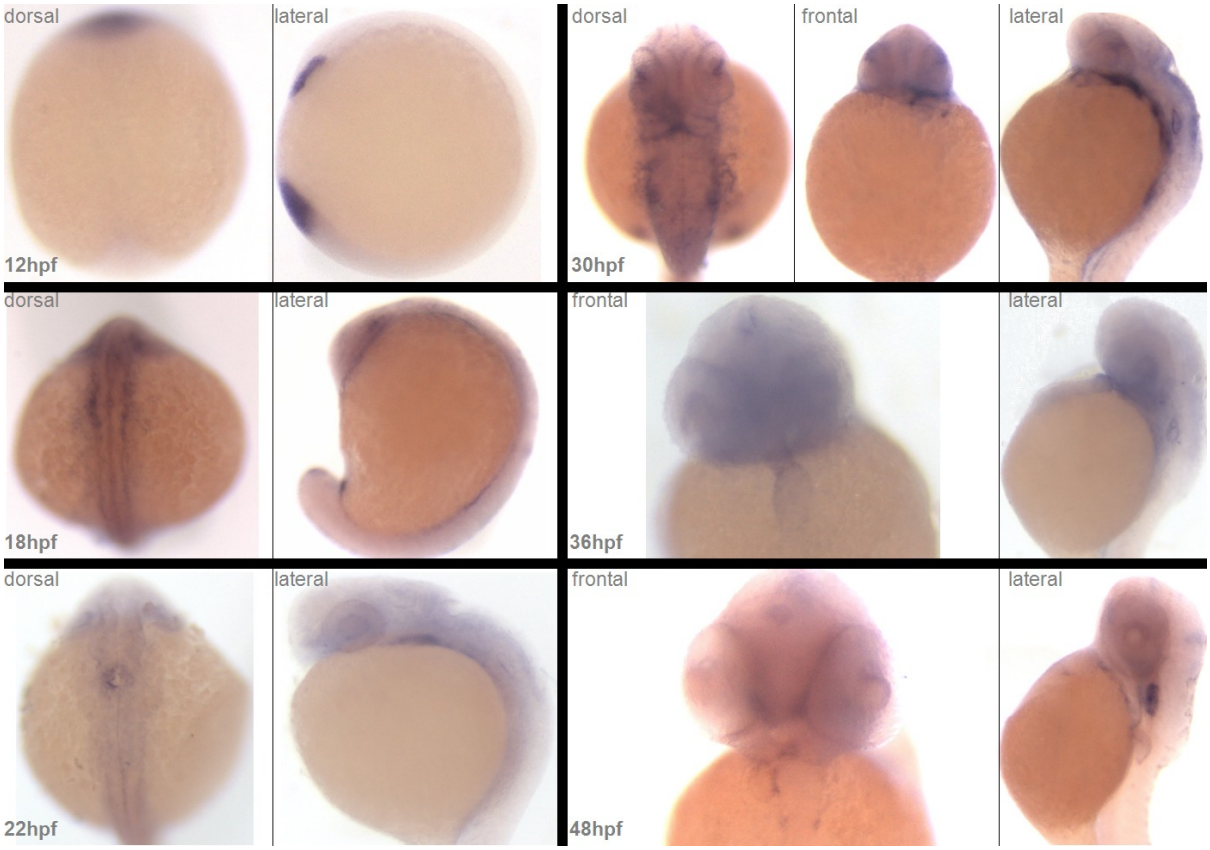


Figure 3: mRNA expression of *bmp4* in the developing zebrafish at different developmental stages. At 18hpf, *bmp4* is expressed in the cardiac progenitor cells that migrate towards the midline. At 30 hpf, bmp4 expressing cells are found in the developing heart as it shifts towards the left. At 48 hpf, *bmp4* is expressed in cells lining the atrioventricular canal.