# An analysis of different ontological approaches to describe renal mutant phenotypes

Kirsty Lee*[1] and Duncan Davidson[2]

[1]School of Informatics, University of Edinburgh, Edinburgh, EH8 9LW and [2]MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh, UK, EH4 2XU

*corresponding author

Email: Kirsty Lee - k.a.lee@sms.ed.ac.uk; Duncan Davidson - Duncan.Davidson@hgu.mrc.ac.uk

## Abstract

Ontologies have increasingly been used in the representation of a variety of biological data. The major alternative ontologies available for mouse phenotype description are the MPO (Mammalian Phenotype Ontology) and PaTO (Phenotype and Trait Ontology). Ontologies have the potential of contributing to the analysis of mutant phenotypes by providing a framework for reasoning. However, any reasoning task will be of limited value if a phenotype ontology cannot represent the majority of phenotypes in publications accurately and in sufficient detail. Therefore, it is important to investigate the accessibility and expressivity of phenotype ontologies, firstly to ensure the scope and consistency of phenotype databases but also as a prerequisite for meaningful automatic reasoning methods. Accessibility in this context is used to refer to the 'ease of use' or how easy it is for researchers to encode phenotype descriptions using the ontology. There have not yet been any published case studies specifically comparing the suitability of current phenotype ontologies for accurately capturing and representing phenotypes using real data sets. This paper incorporates the findings of a 6-month case study which explored potential methods of phenotype description for the EuReGene (European Renal Genome) project. The project uses mouse, rat, zebrafish and *Xenopus* models to examine gene expression patterns and phenotypes relevant to human kidney disease. During the course of the case study, it was possible to visit the participating laboratories which gave a unique and pragmatic insight into how phenotype ontologies can match the requirements of the mouse research community.

## 1 Introduction

The abundance of phenotypic data emerging from mouse mutagenesis screens [1][2] implies a need to describe phenotypes in a way that is amenable to computational comparison. Phenotype comparison is imperative in order to study the underlying genetic mechanisms, and may involve identifying subtle differences between mutant phenotypes. When phenotypic descriptions come in the form of free text, placing lexical and syntactic constraints on them may allow for a more effective comparison.

Recently, ontologies have provided these constraints and have increasingly been used in the representation of a variety of biological data [3]. The major alternative ontologies available for mouse phenotype description are the MPO (Mammalian Phenotype Ontology) [4] and PaTO (Phenotype and Trait Ontology) [5].

Ontologies should be able to contribute to the analysis of mutant phenotypes by providing a framework for reasoning. However, any reasoning task will be of limited value if a phenotype ontology cannot represent the majority of phenotypes in publications accurately and in sufficient detail. Therefore, it is important to investigate the accessibility and expressivity of phenotype ontologies, firstly to ensure the scope and consistency of phenotype databases but also as a prerequisite for meaningful automatic reasoning methods. Accessibility in this context is used to refer to the 'ease of use' or how easy it is for researchers to encode phenotype descriptions using the ontology.

'Ease of use' and expressivity have been highlighted as requirements for the OWL web ontology language, set out in 2004 [6]. Regarding expressivity, the OWL guidelines state that "the language should be as expressive

as possible, so that users can state the kinds of knowledge important to their applications". Regarding the accessibility of the ontology, "the language should provide a low learning barrier and have clear concepts and meaning". Of course, the accessibility of an ontology is dependent on the prior knowledge of the annotator. However, if ontologies are to be used on a large scale, then researchers who may not been directly involved in ontology development could annotate the phenotype data. Therefore, it is desirable to make the annotation process as easy as possible for a non ontology expert.

The tenets set out in the OWL guidelines are equally applicable to phenotype ontologies. However, there have not yet been any published case studies specifically comparing the suitability of current phenotype ontologies for accurately capturing and representing phenotypes using real data sets. For the Gene ontology, Dolan *et al*. (2005) have developed a procedure to address annotation inconsistency using orthologous mouse and human genes [7][7] although there are not yet any similar studies for phenotype ontologies.

This paper will incorporate the findings of a 6-month case study which explored potential methods of phenotype description for the EuReGene project (described further in Section 2) [8]. During the course of the case study, it was possible to visit the participating laboratories which gave a unique and pragmatic insight into how phenotype ontologies can match the requirements of the mouse research community.

## 2 EuReGene Project Overview

The EuReGene (European Renal Genome) project was established in 2005 to bring together expertise from across Europe in order to study kidney function in the normal and diseased states [8]. The project uses mouse, rat, zebrafish and xenopus models to examine gene expression patterns and phenotypes relevant to human kidney disease. Part of the project remit is to allow access to the research outcomes via databases available on the project website. Gene expression and phenotype data are major components of these research outcomes. One important task is to link gene expression and phenotype data which should be achievable by annotating both with a common anatomy ontology. However, as this research is concerned with phenotype description, discussion of gene expression data will be peripheral to the main discussion. Phenotypes are recorded at the various EuReGene research centres and the descriptions are contained in spreadsheets or described using free text. There is an obvious role for phenotype ontologies in standardising the phenotype descriptions. Thus, the EuReGene project provides an opportunity to examine both the accessibility and expressivity of current phenotype ontologies.

## 3 EuReGene Phenotype Data

The EuReGene phenotype data set currently consists of 20 mouse models with 121 phenotype descriptions. The majority of EuReGene phenotype descriptions were for adult mice and relate to kidney physiology; some developmental phenotypes were also described. Phenotype data sheets were submitted by 16 EuReGene partners. The completion of each data sheet required researchers to describe phenotypic characteristics of their model, using free text i.e. without selecting ontology terms. It was important that researchers described the phenotype using free text so that in case of any information loss during the annotation process, the descriptions remained full and accurate. The assays/experimental methods used for phenotype detection were also described on each sheet. Genetic information relating to each animal model was also recorded, such as the targeted gene and the type of genetic manipulation used. Table 1 shows the headings used on each sheet and examples of entries under each heading.

## 4 Encoding EuReGene phenotype data

### *4.1 Ontologies used for encoding phenotypes*

Since the main ontologies available for mouse phenotype descriptions are the MPO (Mammalian Phenotype Ontology) and PaTO (Phenotype and Trait Ontology), these have been used to annotate the EuReGene data.

The MPO has been developed by the Mouse Genome Informatics (MGI) group at the Jackson Laboratory to describe both in-house and external mouse phenotype data from biomedical research literature. The MPO contains 5769[1]

[1] Recorded November 2007

terms for describing phenotypes which are organised into a directed acyclic graph (DAG). The MPO uses atomic terms with each designed to encapsulate a complete phenotype description such as 'polyuria'. In contrast, PaTO is a set of terms to describe phenotypic qualities. (size, shape or colour for example) and is designed to be used in combination with several other ontologies. Using an established ontology such as the MPO holds an advantage over developing an in-house ontology, as the data can be linked with external bioinformatics resources such as phenotype data held at the MGI. For a detailed review of the MPO see [4].

Gene expression data produced by EuReGene members has previously been annotated using EMAP (an anatomical ontology used by the EMAGE database to describe the developing mouse embryo [9][10]). Thus we can associate phenotypes with related gene expression patterns using EMAP.

### *4.2 The Encoding Process*

Each EuReGene phenotype description was annotated (where possible) using the EMAP, MPO and PaTO ontologies. Since PaTO is intended to be used in a compositional framework with other ontologies, there were also three other ontologies used: the Gene Ontology [11], Cell Type Ontology [12] and the ChEBI (Chemicals of Biological Interest) ontology [13]. However, these were used infrequently (see Table 2) and the majority of PaTO qualities were used in conjunction with EMAP anatomical terms. To avoid any annotator bias, each annotation was verified by the researchers involved in creating the mouse models, who had intimate knowledge of the phenotype. Some phenotype descriptions were mapped to two ontology terms but most had a one-to-one mapping. 121 tuples were created for describing EuReGene phenotypes relating to a specific genotype. Table 2 shows the number of terms from each ontology used to describe the EuReGene phenotypes.

## 5 Suitability of ontologies to encode EuReGene descriptions

### *5.1 Suitability of EMAP*

Figure 1 shows that the most commonly annotated EMAP term was 'metanephros'. Annotated EMAP terms were at the organism level (mouse), system level (cardiovascular system), tissue level (renal interstitium group) and organ level (metanephros). Eight phenotype descriptions could not be annotated with an EMAP term (shown in Table 3). The first three of these were examples where a **type** of cell or anatomical part was described rather than a particular named instance. An example is "bone calcification defects". There were four examples of phenotypes at the protein level, for example "normal renin expression". As EMAP is an anatomical ontology, it is unsuitable for annotating these phenotypes. The final example is 'hilar artery' which could not be found in EMAP. However, there was no apparent reason for the absence of hilar artery.

### *5.2 Suitability of MPO*

Figure 2 shows the adequacy of the MPO for describing EuReGene phenotypes. In 45 out of 121 phenotype descriptions an MPO term was found which was either synonymous or an exact replication of the free text phenotype description. There were 44 examples where the MPO could be used but the annotated term was at a higher granularity (less specific) than the free text description. For 33 examples, there was no appropriate MPO term available.

**- MPO Strengths**

Before considering the weaknesses of the MPO, this section considers where the strengths of the MPO lie. Table 4 shows examples where the MPO (but not PaTO) was able to describe the phenotype. The MPO is able to annotate all clinical descriptions in the EuReGene data set, for example 'holoprosencephaly'. Three of the five examples in Table 4 (which are only annotated by the MPO) are clinical descriptions which have corresponding MPO terms but are difficult to describe using PaTO within a compositional framework.

In some cases, PaTO provides an approximate description for clinical descriptions. However, clinical terms in the EuReGene data set are expressed more accurately using the MPO, as exemplified by "hydrops fetalis" (defined by the MPO as "an abnormal accumulation of serous fluid in **fetal tissues**"). An approximate description of "hydrops fetalis" using EMAP and PaTO is 'mouse' + 'edematous'. However, the PaTO description is less accurate than the MPO since fetal tissue is not incorporated in the description. There are further examples of clinical terms which can be

annotated with both ontologies but are much more intuitively annotated using the MPO, such as "hypercalciuria". These examples are discussed in relation to the suitability of PaTO in Section 5.3.

**- MPO Weaknesses**

In order to identify the weaknesses of the MPO, it is necessary to study examples where the MPO was not able to describe the phenotype. However before these examples are discussed, it is also important to consider examples where the MPO annotation resulted in a less specific phenotype description. These examples are shown in Table 5 and correspond to the third category shown previously in Figure 2. For each example, the text in bold shows where the specificity has been lost. The curly brackets after the MPO term in Table 5 point to the general type of term where the specificity was lost. In 6 examples, the anatomical specificity is insufficient and in 7 examples the specificity of the quality (e.g. tortuosity) has been lost.

Although the MPO may not aim to describe phenotypes at a detailed cellular level, in an example such as "low molecular weight proteinuria" there is information lost after annotation with the MPO term 'proteinuria'. The presence of "low molecular" in this example allows useful distinctions to be made regarding the underlying kidney filtration processes.

Many of the EuReGene phenotypes are biochemical measurements made in the blood and urine. The two main functions of a kidney are to filter the blood and excrete waste products in the form of urine. Thus the disruption of normal kidney function can be identified by examining the concentration of various substances, such as amino acid, in the blood and urine. The clinical term 'aminoaciduria' is commonly used to describe an increase in the concentration of amino acid present in the urine. Equivalent terms for calcium and protein concentrations are 'hypercalciuria' and 'proteinuria'. Similar terms for describing ion concentrations in the blood are 'hypercalcemia' (high calcium) and 'hypokalemia' (low potassium). The MPO incorporates these and other similar clinical terms and thus minimal translation from free text is required. In contrast, using PaTO to annotate this category of phenotypes is slightly more cumbersome and is described below when considering PaTO annotation.

Having considered examples where the MPO description was less specific, it is now appropriate to consider examples where no MPO term was suitable (corresponding to the fourth category in Figure 2.). In order to make general inferences which could apply to phenotypes outwith EuReGene, the reasons for non-annotation have been categorised as follows:

1. **The MPO could not describe a normal phenotype referring to a particular process or anatomical part**

   Example: "normal electrolyte levels in the blood"

2. **The MPO could not describe the absence of a particular anatomical part or process**

   Examples: "no apoptosis", "no haemorrhage"

3. **The granularity of the phenotype description meant that no MPO term was available.**

   Examples: "SorLA protein upregulation", "Renal Fanconi syndrome", "cardiac conotruncal defects"

4. **No reason could be established and additional terms should be added to the MPO**

Currently, the MPO terms available for describing normal phenotypes are 'normal phenotype' and 'no abnormal phenotype detected'. However, there is no MPO term to describe a normal phenotype with reference to a particular process or anatomical part such as 'normal electrolyte levels in the blood'. As a result, MPO annotators use terms which are also used to specify the anatomy without any abnormality. For example 'muscle phenotype' is used to annotate the free text description "no obvious muscle abnormalities". A similar example is the free text description "at E18.5 there are no discernable gross abnormalities in the kidneys" which has been annotated with 'renal/urinary system phenotype'. These examples may cause problems for any future reasoning as 'muscle phenotype' is being used to describe both normal phenotypes (as shown above) and abnormal phenotypes. The same is true for 'renal/urinary system phenotype' and probably other similar terms.

Because the MPO must add a term to describe the absence of every process or anatomical part, it is less likely to have a term linked to every process mentioned in GO or every anatomical part included in EMAP. Thus, the extensibility of the MPO compared to PaTO means that the former is less likely to be able to describe absent entities.

Seven examples where the granularity was a problem were related to a protein. However, there were also two examples where the granularity of the free text description was too general to be described in the MPO. These were "Renal Fanconi syndrome" and "cardiac conotruncal defects".

### *5.3 Suitability of PaTO*

In comparison with the MPO, there are marginally more examples which can be annotated with PaTO (shown in Figure 3). Due to the compositional nature of the ontology, there were no examples where the ontological description exactly matched the free text description, as was the case with the MPO (in 14% of examples). However, a much higher proportion (63%) of the examples are synonymous with the free text descriptions, compared with 23% for the MPO.

**- PaTO strengths**

In order to identify the strengths of PaTO, it is useful to examine the phenotype descriptions which could only be annotated using PaTO. Table 6 contains examples which were annotated with PaTO but not the MPO. Examples where no appropriate entity term existed (although PaTO quality terms did) are also included in Table 6.

Absent phenotypes

Examples 1-3 in Table 6 describe the **absence** of either a process (apoptosis) or anatomical part (nephron, renal vesicle). These are easier to annotate using PaTO since 'absent' can be applied to any entity which is available in EMAP/GO or other external ontology. Example 4 which describes the absence of a haemorrhage is slightly more difficult as an entity for haemorrhage could not be found.

General phenotypes

PaTO can describe the general dysfunction of any continuant entity (e.g. anatomical part) using the quality 'decreased functionality'. 'Functionality' can be associated with a term from an external ontology such as EMAP. As the EMAP anatomy ontology tends to have a higher granularity than the anatomy terms in the MPO, "generalized proximal tubule dysfunction" (Example 5) can be described using PaTO but not the MPO. Generally, PaTO is more suited to describing phenotypes that are general abnormalities as it is not as constrained by the granularity of the anatomical/process description. The entities in these cases are only constrained by the terms available in other OBO ontologies rather than the coverage of the MPO. Since these other ontologies tend to contain more specific anatomical parts (for example EMAP) and processes (GO), a higher specificity of phenotype description is achieved.

Process phenotypes

Examples 6-7 show an advantage of the compositional approach by allowing the description of a process (endocytosis) within a specific anatomical part (renal proximal tubule). PaTO is able to describe many different features of processes (impaired and abolished in these examples) whilst also describing where they occur. Consistent with the earlier examples, provided external ontologies can match the desired specificity, PaTO is more flexible and able to cope with many different processes in combination with various permutations of **how** and **where** they are affected.

Other examples

Unlike the earlier examples which demonstrate the advantages of the compositional nature of PaTO, Examples 8-13 do not reflect a particular design advantage of one ontology over the other. Instead, they reflect cases where appropriate terms were available in PaTO but not in the MPO, for example 'drug_response' which is used to describe Example 9.

**- PaTO weaknesses**

Table 7 shows examples where the PaTO description was less specific than the free text description. These examples correspond with the second category shown in Figure 3. In several examples, supplementing PaTO with additional terms would remove this information

loss. In the first four examples, the specificity of the quality is lost, for example the quality term 'tortuosity' should be added to describe "increase in vessel tortuosity". Where PaTO could not be used to describe a phenotype, the reasons were categorised as shown below. Examples are given for each category, except where there was no obvious reason and thus no commonality between the descriptions.

1. **PaTO could not annotate because the phenotype was a clinical description.**

   Examples: "holoprosencephaly"
   "renal tubular acidosis"
   "cardiac conotruncal defects"

2. **PaTO could not annotate because the phenotype described protein or mRNA expression.**

   Examples: "overexpression of Cd2ap and Nphs2 mRNA"
   "normal angiotensinogen expression" (also an example of category 3 below)
   "normal renin expression" (also an example of category 3 below)

3. **PaTO could not represent a normal phenotype related to a specific entity.**

   Example: "normal urine concentrating ability"

4. **No obvious reason**

It is often significant if a normal phenotypic result appears on a mutant background since it may show dependence of the phenotype on environmental factors or development stage. Or in a double mutant the effects of a second gene may compensate for the actions of the first producing an apparently normal phenotype. A normal phenotype may also signify that the mutant allele is not involved in the biological process or function which is being studied. Several normal phenotypes exist in the EuReGene data set, for example "normal urine concentrating ability" and "no abnormal electrolyte concentrations in blood". PaTO was able to describe a higher proportion of normal phenotypes than the MPO. There were 6 examples where the MPO could not describe the normal phenotype and 4 examples where PaTO could not. PaTO is able to describe normal phenotypes using the term 'normal' which can be applied to any entity, whereas we have seen that the MPO does not contain any terms for relating a normal phenotype to a specific entity.

Although PaTO can describe any process, cell or anatomical part using 'abnormal', there remains a difficulty with representing many normal qualities. In the EuReGene data set, normal concentration phenotypes illustrated the difficulty. However, this could equally apply to qualities such as 'amount', 'temperature', 'size'; indeed any other quality which can be increased/decreased (currently there are 72 of these) can also be normal. Thus, at least 72 terms could be added to the PaTO ontology which would allow an increased range of normal phenotypes to be expressed. A similar situation arises with abnormal phenotypes where the increase or decrease in a quality may be unspecified.

## 6 Discussion

### *6.1 EMAP Expressivity*

EMAP appeared to be capable of annotating the majority of EuReGene phenotypes. Protein expression phenotypes and those where a general type of entity was included (e.g. bone) were the only problematic descriptions. The only anatomical phenotype term unavailable was 'hilar artery' which should be added to the EMAP ontology.

### *6.2 Expressivity using a compositional framework*

PaTO allows the flexibility of combining any anatomical, process or cell entity with any phenotypic quality which resulted in a higher proportion of annotation within the EuReGene data set. However, with PaTO annotation there is a danger that the power of combining multiple ontologies will be lost if the entity essentially describes the phenotype and the quality is 'abnormal'.

PaTO offers flexibility by combining terms from multiple ontologies. There were 39 examples in the EuReGene dataset where two entities were required to complete the full phenotype description using PaTO qualities. 29 were phenotypes describing the concentration of an entity, 8 described processes, (3 absorption, 4 endocytosis, 1 apoptosis), 1 related to a drug response and the final was a general cell type.

Many phenotypes involve the alteration of a process within a specific anatomical context. For example, "impaired proximal tubular endocytosis" describes the impairment of the process 'endocytosis' but also specifies that this occurs in the 'proximal tubule'. PaTO is well designed for describing the temporal features of a process using terms such as 'arrested' or 'delayed'. However, there is an additional requirement for a mechanism to describe **where** the process is affected. In practice, annotators using PaTO often use several 'entity' terms from various ontologies to complete the full phenotype description. In the example above we could use the GO process term 'endocytosis' supplemented with the anatomy term 'proximal tubule' which can then be used with a PaTO quality (impaired) to form the full phenotype description. This process has been termed "post-composition". However, post-composition in this context could be fairly arbitrary.

Descriptions are post-composed at the point of annotation and term associations are not necessarily approved by the ontology developers at this point. This differs from the MPO where although terms may be more constrained, they are fixed in the ontology. By allowing the flexibility of post-composing descriptions *ad hoc*, the rigour of the ontology may be compromised. Therefore, there should be a more formal structure for post-composition to prevent many synonymous phenotypes being annotated with slightly different combinations of ontology terms.

### *6.3 Usability*

The MPO and PaTO are based on two distinct approaches to phenotype description. The evolution of MPO has been driven by the phenotype descriptions appearing in journal articles resulting in many MPO terms closely resembling free text descriptions. From an annotation perspective, it is advantageous to use ontology terms which mirror descriptions already established in the field such as 'renal tubular acidosis'.

In the majority of examples, it was simpler to translate the descriptions provided by EuReGene researchers to MPO terms. This became apparent when discussing appropriate ontology terms with EuReGene partners and is confirmed by the higher proportion of free text descriptions which are exactly the same as the MPO term, shown in Figure 2. However, this does not take into consideration the possible downstream applications of the phenotype ontology. It may be appropriate to use MPO for annotation and then link this to an underlying PaTO description which can be used for reasoning. This would rely on an accurate translation from the MPO to PaTO. Efforts towards formalising this translation have begun within the National Centre for Biomedical Ontology [14].

It is important to enable collaboration between biology researchers and phenotype annotators. Developing an expertly curated phenotype data set relies on the insight of researchers who have in-depth knowledge of each phenotype. To ensure consistent use of ontology terms, an independent researcher with knowledge of phenotype ontologies is ideal. Other possibilities for ensuring consistency would be to develop detailed, publicly available guidelines on annotation and/or an online submission system which suggests previously curated ontology terms.

## Acknowledgements

EuReGene (www.euregene.org) is an Integrated Project funded by the
EC as part of the Framework program 6 (FP6 005085).

## 7 References


[1] The European Mouse Mutagenesis Consortium: The European dimension for the mouse genome mutagenesis program. *Nature Genetics* 2004, 36(9): 925-7.

[2] The International Mouse Knockout Consortium: A Mouse for all reasons. *Cell* 2007, 128: 9-13.

[3] Open Biomedical Ontologies (OBO) [http://www.obofoundry.org/]

[4] Smith CL, Goldsmith C-AW, Eppig JT: The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology* 2005, 6(1): R7

[5] PaTO (Phenotype and Trait Ontology) [http://www.obofoundry.org/cgi-bin/detail.cgi?id=quality]

[6] OWL Web Ontology Use Cases and Requirements (2004)

[http://www.w3.org/TR/webont-req]

[7] Dolan ME, Ni L, Camon E, Blake JA: A procedure for assessing GO annotation consistency. *Bioinformatics* 2005, 21(Suppl 1):i136-43

[8] Willnow, T. et al. The European Renal Genome Project: An Integrated Approach Towards Understanding the Genetics of Kidney Development and Disease. *Organogenesis* 2005, 2:2: 42-47

[9] Christiansen JH, Yang Y, Venkataraman S, Richardson L, Stevenson P, Burton N, Baldock RA & Davidson DR: EMAGE: A spatial database of gene expression patterns during mouse embryo development. *Nucleic Acids Res*. 2006, 34: D637-41

[10] Baldock RA, Bard JB, Burger A, Burton N, Christiansen J, Feng G, Hill B, Houghton D, Kaufman M, Rao J, Sharpe J, Ross A, Stevenson P, Venkataraman S, Waterhouse A, Yang Y, Davidson DR.: EMAP and EMAGE: a framework for understanding spatially organized data. *Neuroinformatics* 2003, 1(4): 309-25

[11] Harris, M.A. et al. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 2004, 32 (Database issue): D258-261

[12] Bard, J. et. al.: An Ontology for Cell Types. *Genome Biology* 2005 6:R21

[13] Chemicals of Biological Interest (ChEBI) [www.ebi.ac.uk/chebi/]

[14] National Centre for Biomedical Ontology [http://www.bioontology.org/]

## Tables

Table 1. EuReGene phenotype description headings with examples

| **PHENOTYPE DESCRIPTION SHEET HEADING** | **EXAMPLE** |
|---|---|
| Gene targeted | Clcn5 |
| Type of genetic manipulation | Knock out |
| Spatial and structural information | Proximal tubule<br>(subcellular localization to endosomes) |
| Major phenotypes, in relation with structural defects | • Low molecular weight proteinuria<br>• Generalized aminoaciduria<br>• Glycosuria<br>• Hypercalciuria- renal calcium deposits<br>• Increased bone turnover<br>• Impaired proximal tubular endocytosis |
| Assays used to determine the phenotypes | Standard chemistry, immunohistochemistry |
| Biological data available on individual or pool basis? | Pool basis |
| Publications using the mouse | Pubmed ID : 11115837 |

Table 2. The number of terms annotated to EuReGene phenotypes for each ontology. GO, cell type and ChEBI ontologies were used in conjunction with PaTO terms where an anatomical term was not appropriate. The number of phenotype descriptions which could not be annotated is shown in brackets. Since the final three ontologies were only used in conjunction with PaTO qualities, numbers in brackets are not included for these.

| Ontology | NUMBER OF DISTINCT TERMS USED TO ANNOTATE PHENOTYPES |
|---|---|
| Edinburgh Mouse Atlas Project (EMAP) Ontology | 31 (8) |
| Mammalian Phenotype Ontology (MPO) | 53 (32) |
| Phenotype and Trait Ontology (PaTO) | 32 (23) |
| Gene Ontology (GO) | 6 |
| Chemical Entities of Biological Interest (ChEBI) | 2 |
| Cell type Ontology | 1 |

Table 3. Examples of free text phenotype descriptions where no EMAP term was available.

| **FREE TEXT PHENOTYPE DESCRIPTION** | **REASON WHY EMAP WAS UNSUITABLE** |
|---|---|
| bone calcification defects | Anatomical term was a general type |
| increased bone resistance | Anatomical term was a general type |
| stromal cell defect (immediately adjacent to renal pelvis region) | Anatomical term was a general type |
| normal angiotensin 1 | Protein level could not be described |
| normal angiotensinogen expression | Protein level could not be described |
| normal renin expression | Protein level could not be described |
| enhanced processing of amyloid precursor protein (APP) to amyloid in neurons in the brain | Protein level could not be described |
| hilar artery calcification | Missing term – unknown reason |

Table 4. EuReGene phenotype descriptions (free text) which were annotated by only the MPO. Exact match and synonymous MPO annotations are included (first 2 categories shown in Figure 2). Definitions are included for each MPO term.

| FREE TEXT PHENOTYPE DESCRIPTION | MPO TERM | MPO TERM DEFINITION |
|---|---|---|
| distal renal tubular acidosis | renal tubular acidosis | a clinical syndrome characterized by the inability to acidify urine |
| holoprosencephaly | holoprosencephaly | presence of a single forebrain hemisphere or lobe; often accompanied by a deficit in median facial development |
| polyhydramnios | polyhydramnios | abnormally high amniotic fluid volume; may result from maternal diabetes, chromosomal abnormalities or other congenital abnormalities |
| high lethality | premature death/postnatal lethality | death after weaning age, but before the normal life span/premature death anytime after postnatal day 1 to weaning age |
| concentration defect | abnormal urine osmolality | changes in the concentration of ions in the urine compared to the normal state |

Table 5. Examples where the MPO term was less specific than the free text description. The curly brackets point to the category of term which could not be expressed with sufficient specificity.

| Free text description | MPO term | |
|---|---|---|
| increase in vessel **number**<br>increase in vessel **tortuosity** | abnormal vasculature | |
| no obvious **histological** lesions noted (glomerulus) | no abnormal phenotype detected | quality |
| renal iron **deposits** | abnormal kidney iron level | |
| tubular **atrophy** | renal tubular necrosis | |
| podocyte **hypertrophy**<br>podocyte **vacuolization** | abnormal podocyte | |
| altered vascular activity/abnormal **calcium signalling** | abnormal vascular endothelial cell physiology | process |
| bone **calcification** defects | abnormal bone mineralization | |
| **rare crescent formation** | abnormal renal glomerulus morphology | |
| **peritubular capillary** regression | abnormal vascular regression | |
| **arcuate artery** calcification | arterial calcification | anatomical |
| **focal** and **segmental** glomerulosclerosis | glomerulosclerosis | |
| calcification in the **renal papilla** | kidney calcification | |
| increased **plasma vitamin D3** | abnormal vitamin level | |
| urinary excretion of **vitamin A** bound to retinal binding protein | abnormal retinol metabolism | vitamin |
| **interstitial** haemorrhage | kidney hemorrhage | |
| actin-expressing **smooth muscle cells** fail to differentiate | abnormal cell differentiation | cellular |
| **increased** mesangial **matrix** | abnormal mesangial cell | |
| **low molecular weight** proteinuria | proteinuria | protein |

Table 6. EuReGene phenotype descriptions (free text) which were annotated by PaTO only. (Corresponds with first category shown in Figure 3).

| **EXAMPLE NO.** | **FREE TEXT DESCRIPTIONS WHERE ONLY PATO COULD ANNOTATE** |
|---|---|
| 1 | nephrons fail to develop/lack of nephrons |
| 2 | no apoptosis |
| 3 | renal vesicles do not form |
| 4 | no haemorrhage |
| 5 | generalized proximal tubule dysfunction (Renal Fanconi syndrome) |
| 6 | loss of endocytic activity in the renal proximal tubules * |
| 7 | impaired proximal tubular endocytosis * |
| 8 | stromal cell defect (immediately adjacent to renal pelvis region) |
| 9 | impaired response to loop and thiazide diuretics |
| 10 | vascular fragility |
| 11 | decreased lithium clearance |
| 12 | increased chloride excretion * |
| 13 | increased aldosteronuria |

***** denotes that 2 examples of this free text description were present in the EuReGene data set

Table 7. Examples of phenotype descriptions where using PaTO resulted in a loss of specificity. The bold terms indicate where the specificity has been lost.

| Example no. | Free text phenotype description | PaTO description | | |
|---|---|---|---|---|
| | | **Entity** | **PaTO parent** | **PaTO child** |
| 1 | increase in vessel **tortuosity** | renal cortical arterial system | structure | disorganized |
| 2 | peritubular capillary **regression*** | renal cortical capillary | relative quantity | decreased |
| 3 | no obvious **histologic** lesions noted (glomerulus) | glomerular tuft | deviation (from normal) | normal |
| 4 | **perinatal /postnatal** lethality* | mouse | viability | dead |
| 5 | **focal** and **segmental** glomerulosclerosis ** | glomerular tuft | structure | collapsed |
| 6 | hydrops **fetalis** | embryo | structure | edematous |
| 7 | **microvillus** formation | visceral epithelium | structure | degenerate |
| 8 | **rare crescent formation** | Bowman's capsule | structure | hyperplastic |
| 9 | **foot process** effacement* | visceral epithelium | deviation (from normal) | abnormal |
| 10 | mesangio**lysis**** | glomerular mesangium | structure | degenerate |

* denotes that 2 examples of this free text description were present in the EuReGene data set.

** denotes that 3 examples of this free text description were present in the EuReGene data set.

## Figures

Figure 1. Frequency of EMAP terms in the EuReGene phenotype data set. Terms are listed in ascending order of frequency.

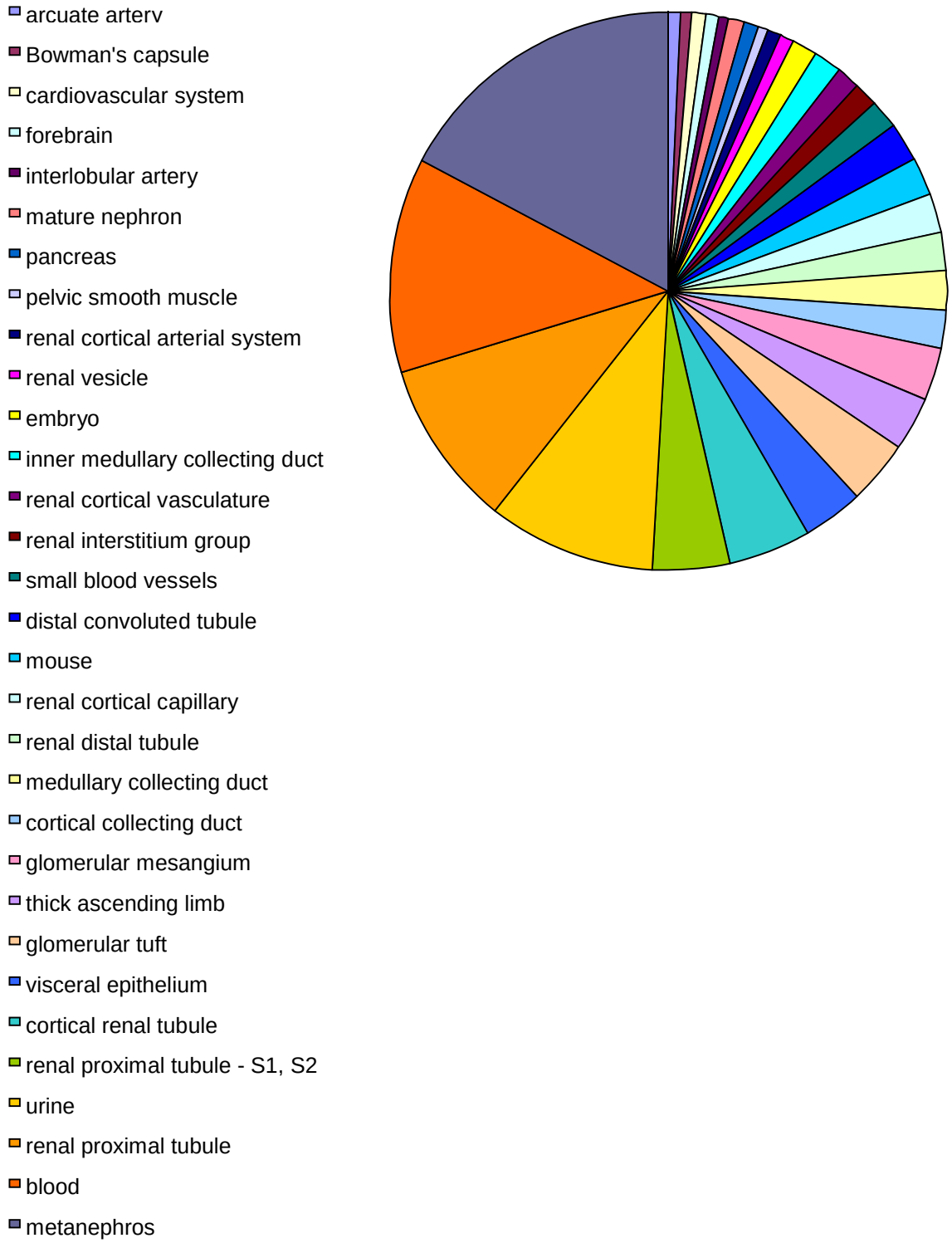
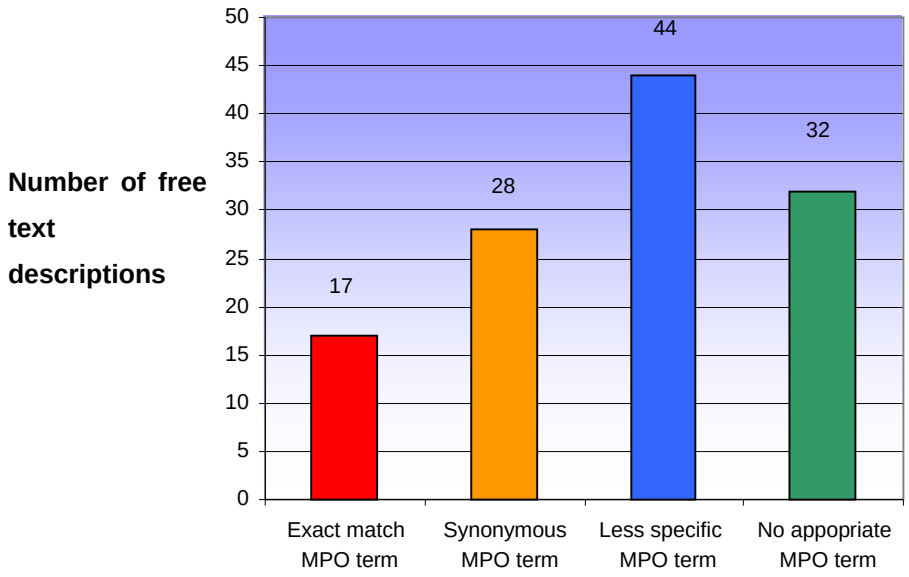


Figure 2. Frequency of MPO annotation for 121 EuReGene phenotypes (originally described using free text). If a free text phenotype description is repeated (e.g. by different labs), both are counted individually.




Key

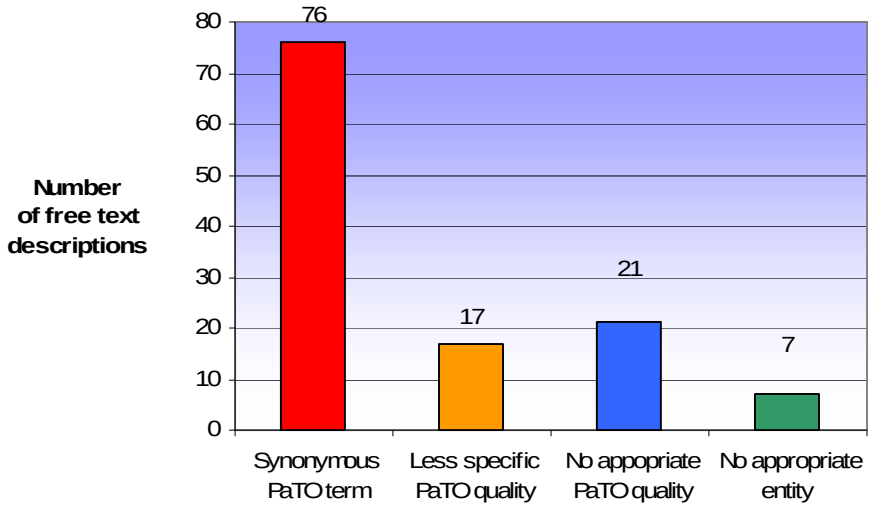
Exact match MPO term: the annotated MPO term was exactly the same as the free text phenotype description

Synonymous MPO term: the annotated MPO term was synonymous with the free text phenotype description

Less specific MPO term: the annotated MPO term was less specific than the free text description

No appropriate MPO term: there was no appropriate MPO term to describe the phenotype

Figure 3. Frequency of PaTO annotation for 121 EuReGene phenotypes (originally described as free text). If a free text phenotype description is repeated (e.g. by different labs), both examples are counted individually.




Key:

Synonymous PaTO term: the annotated PaTO term was synonymous with the free text phenotype description

Less specific PaTO term: the annotated PaTO term was less specific than the free text description

No appropriate PaTO term: there was no appropriate PaTO term to describe the phenotype

No appropriate entity: there was no appropriate term for describing the entity part of the PATO description, for example if there was no appropriate GO or cell type term