DMD Fast Forward. Published on April 16, 2012 as doi:10.1124/dmd.112.045062

Probabilistic orthology analysis of the ABC transporters: Implications for the development of multiple drug resistance phenotype

Ciaran Fisher, Tanya Coleman and Nick Plant

Centre for Toxicology, Faculty of Health and Medical Sciences, University of Surrey, Guildford, Surrey GU2 7XH. CF and NP

Clinical Pharmacology and DMPK, AstraZeneca Clinical Development, Alderley Park, Macclesfield, Cheshire, SK10 4TJ. TC

Probabilistic orthology analysis of the ABC transporters

Author for correspondence

Dr Nick Plant

Centre for Toxicology,

Faculty of Health and Medical Sciences,

University of Surrey, Guildford,

GU2 7XH, UK

Tel: +44 (0)1483 686412

Fax: +44 (0)1483 686401

Email: N.Plant@Surrey.ac.uk

Number of Text Pages: 21

Number of Tables: 1

Number of Figures: 3

Number of References: 47

Words in Abstract: 213

Words in Introduction: 743

Words in Discussion: 845

Abbreviations: ABC, ATP Binding Cassette; ABD, ATP Binding Domain; ADME, Absorption, distribution, metabolism and excretion; MDR, Multiple Drug Resistance; MPR, Most Parsimonious Reconciliation; PK, pharmacokinetics; TMD, Transmembrane domain; UPMGA, Unpaired Grouping Method with Arithmetic Mean

DMD Fast Forward. Published on April 16, 2012 as DOI: 10.1124/dmd.112.045062
This article has not been copyedited and formatted. The final version may differ from this version.

Abstract

Drug transporters are rapidly becoming recognised as central to determining a chemical's fate within the body. This action is a double-edged sword, protecting the body from toxicants, but also potentially leading to reduced clinical efficacy of drugs through multiple drug resistance phenotype. To examine the inter-relationship of this super-family we have constructed phylogenetic trees over an extended evolutionary distance representing each of the seven sub-families. In addition, using protein sequences from species important in the design and evaluation of novel chemicals, namely human, macaque, rat, mouse and dog, we have undertaken probabilistic orthology analysis to examine speciation probabilities within this phylogeny. This data allows us to accurately predict orthologous sequences across these species, an important confirmatory step with implications for cross-species extrapolation of data during drug safety testing. Finally, we present the first complete phylogeny for sub-families within humans constructed utilising the entire coding sequences, at both the DNA and protein levels. We demonstrate for the first time that genes associated with the multiple drug resistance phenotype cluster separately from other genes within the same sub-family, suggestive of a conserved, fundamental, difference in these proteins. Such work may help guide future studies on the mechanisms underlying multiple drug resistance, as well as the development of novel therapeutic approaches to mitigate against its development.

DMD Fast Forward. Published on April 16, 2012 as DOI: 10.1124/dmd.112.045062
This article has not been copyedited and formatted. The final version may differ from this version.

Introduction

The ATP-binding cassette (ABC) genes comprise a super-family with representatives found in all characterised eukaryotic and prokaryotes; indeed, this superfamily encodes approximately 5% of the *Escherichia coli* genome (Fath and Kolter, 1993; Davidson and Chen, 2004). The majority of ABC genes encode membrane bound transport protein, which act to move polar molecules across the non-polar lipid membrane, utilising the hydrolysis of ATP. As such, these transporters play an important role in the absorption, distribution, metabolism and excretion (ADME) of chemicals (Glavinas et al., 2004). In prokaryotes, ABC transporters may act as both importer and exporter proteins (Fath and Kolter, 1993; Davidson and Chen, 2004). By contrast, in eukaryotes, these proteins act solely as export transporters, and this represents an important functional breakpoint within the super-family. Such efflux is central to the removal of potentially harmful chemicals from cell systems; an action that undoubtedly underlies the biological survival advantage conferred by these proteins and explains their conservation across evolutionary time (Dean and Annilo, 2005).

Although the ability to rapidly eliminate potentially harmful chemicals has obvious survival advantages, it also represents a challenge during long-term chemotherapy. Expression levels of a number of ABC transporters has been shown to contribute towards the development of multidrug resistance (MDR) phenotype, whereby the ADME of administered chemicals is altered, usually resulting in altered pharmacokinetics (PK) and reduced clinical efficacy. MDR has been shown to have a negative impact on the treatment of a number of disease states, including cancer (Deeley et al., 2006; Gillet et al., 2007). Much work has thus been undertaken to understand the molecular mechanisms underlying MDR, and how this can be mitigated during long-term chemotherapy (Coley, 2008).However, translation of these mitigation strategies to the clinic has generally been poor, and MDR still represents a

significant hurdle to successful chronic chemotherapy regimens (Coley, 2008; Tiwari et al., 2011).

Previous studies on the evolution of the ABC genes have not included all the species utilised in the pre-clinical testing of novel drugs; such a comprehensive analysis would be important for the robust extrapolation of data from preclinical test species to humans. In addition, phylogenetic analysis has often been restricted to only fragments of the total coding sequence, such as the ATP-binding domain, which is likely to be the least variable domain (Dean and Allikmets, 1995; Saier and Paulsen, 2001; Dean and Annilo, 2005), and has relied on distance-based methodologies which are generally accepted to not produce the most robust phylogenetic relationships across super-families (Koski and Golding, 2001).

In the current study, we have used protein alignments of all members of the ABC drug transporter family in humans and a number of important model animals for the testing of novel chemicals to undertake probabilistic orthology analysis. This allows the robust assignment of ortho- or paralogue status to protein pairs, including probability values, thus providing important information for extrapolation of effects between species. In addition, we have expanded on previous phylogenetic studies by using full ABC nucleotide and protein sequences across a range of evolutionarily diverse species. Such an approach not only separates ABC sub-families correctly, but also clusters those ABC transporters associated with MDR, suggesting that this is a specifically evolved design function.

DMD Fast Forward. Published on April 16, 2012 as DOI: 10.1124/dmd.112.045062
This article has not been copyedited and formatted. The final version may differ from this version.

Materials and Methods

Sequence identification and alignment of ABC genes

The Reference Sequences (RefSeq) for all 48 members of the human ABC super-family were identified (Supplemental Table 1) along with their common alternate names and GenBank accession number. Sub-families A-D and G represent those sequences that encode membrane-bound drug transport proteins, whereas sub-families E and F encode ATP binding cassette proteins that do not play a role in drug transport. Using these human sequences as the query, a cross-species megaBLAST was undertaken to identify similar sequences in ten other species (Altschul et al., 1990; Zhang et al., 2000). Where multiple sequences against a single query were identified within a species, all sequences were taken forward, with duplicates trimmed when identified.

Multiple sequence alignments was performed for each sub-family using ClustalX, with an additional alignment being performed solely for the human sequences (Higgins and Sharp, 1988; Thompson et al., 1997): For each sub-family, an *E. coli* ABC gene sequence was included, as the root for the subsequent phylogenetic tree. All alignments were assessed manually to ensure gap insertions were sensible, with subsequent phylogenetic analysis also automatically removing regions containing significant gaps.

Phylogenetic analysis of ABC sub-families

Distance matrices were produced from the ClustalX output using dnadist, including 100 replicates for bootstrap analysis (Felsenstein, 1997). For individual sub-family analysis, these

matrices were analysed with Neighbor using an unpaired grouping method with arithmetic mean (UPGMA) to produce a consensus tree using the majority rule (Felsenstein, 1997).

For analysis of all sub-families within humans, ClustalX-generated alignments were analysed with Phyml, as this is more computationally efficient for larger analysis (Felenstein, 1989; Guindon and Gascuel, 2003). One hundred phylogenetic trees were constructed using a maximum-likelihood approach, and Consense was used to generate a consensus tree from the bootstrapped data (Felenstein, 1989).

Probabilistic orthology analysis of ABC transporter proteins from human, mouse, rat, dog and macaque

Probabilistic orthology analysis is a reconciliation-based orthology methodology that utilises the probabilistic gene evolution model described by Arvestad et al. (Arvestad et al., 2003). The output determines the probability that any given divergence within a phylogenetic tree is the result of speciation as opposed to duplication. This allows the robust assignment of orthologues and paralogues within a tree. Protein sequences for each ABC drug transporter from human, dog, rat, mouse and macaque, were aligned using ClustalX2 (v2.0.12), and ProTest used to determined that LG was the optimum amino acid replacement model for accurate phylogeny resolution (Abascal et al., 2005; Le and Gascuel, 2008). A maximum-likelihood phylogenetic tree was generated with Phyml, which was the input for the probabilistic orthology analysis. In addition to a phylogenetic relationship of genes, a phylogenetic relationship of species is also required for probabilistic orthology analysis, and this was constructed using species divergence times taken from Kumar and Hedges (Kumar and Hedges, 1998) and Jacobs and Downs (Jacobs and Downs, 1994). Probabalisitic orthology analysis was undertaken using PrimeGEM, as described in Sennblad and Lagergren

(Sennblad and Lagergren, 2009), using 10000 MCMC iterations, MCMC-estimated duplication and loss rates, and with an output of posterior orthology probabilities. This output was analysed using the MCMC_analysis perl script available from http://prime.sbc.su.se/primeGEM/downloads/perl/mcmc_analysis.

DMD Fast Forward. Published on April 16, 2012 as DOI: 10.1124/dmd.112.045062
This article has not been copyedited and formatted. The final version may differ from this version.

Results

Phylogenetic analysis of ABC sub-families

Using a distance-based method we were able to generate nucleotide-level phylogenetic trees for each sub-family (Supplemental Figures 1-5). Table 1 represents a summary of these data, describing the total number of ABC genes identified in each species, plus the number of sequences for each species that could be clearly demonstrated to lie within a single sub-family through phylogenetic analysis.

Using distance-based methods it is only appropriate to assign individual sequences within a sub-family, and not to predict orthologues. Assignment of orthologues based purely upon phylogenetic trees derived by most parsimonious reconciliation (MPR) has been demonstrated to have poor predictive value (Koski and Golding, 2001). A more appropriate analysis is probabilistic orthology analysis (Sennblad and Lagergren, 2009). In this method, gene evolution is set within the context of a species evolution tree and modelled under variable gene birth-death rate parameters (Rannala and Yang, 2007).

For each human ABC protein, we present the most probable orthologue for macaque, rat, mouse and dog (Supplemental Table 2), along with the probability score for that match. In general, given nomenclature for each protein is consistent with the indicated analysis; for example, the protein named ABCA6 in humans and mice are demonstrated to be the result of a speciation event, and hence be orthologues, with a probability of over 99%. However, some orthologue assignments are not fully supported by the probabilistic orthology analysis: For example, in the case of ABCD2, probabilistic orthology analysis assigns high probabilities to the sequence being an orthologue of either ABCD2 or ABCD3. In such cases, sequences are tentatively assigned as orthologues to the human sequence for which there is the higher

probability, but alternate high probability matches are noted within the table (Supplemental Table 2) and phylogeny (Supplemental Figure S6). Finally, we are able to expand the current knowledge base, assigning orthologue status to several 'orphan' sequences, particularly from the Macaque, which has been relatively poorly investigated until now.

In addition, probabilistic orthology analysis provides an estimate of the most probable gene loss/duplication rates that would result in the given phylogeny. For each combination of duplication and loss rates examined within the analysis, the probabilistic likelihood is determined, producing the 3-dimensional Gaussian distribution seen in Figure 1. The maximum posterior probability of the analysis represents the most likely duplication and loss rate estimates, which were $0.0095Myr^{-1}$ and $0.0122Myr^{-1}$, respectively. It is important to note, that these estimates are relevant for only the species set examined (human, rat, mouse, dog, macaque), with duplications and loss rates often being considerably different over larger evolutionary distances. .

Phylogenetic analysis of the ABC super-family in humans

Following examination of phylogenetic relationships within sub-families across a number of animal species relevant to pre-clinical chemical testing, we next examined the relationship between human sub-families. Phyml was used to confer a logical consensus tree with acceptable bootstrap values at major nodes (Figure 2). This analysis was able to successfully resolve the sub-families, and provide further insight into the evolution of this super-family. Unsurprisingly, sub-families E and F, which lack a transmembrane domain (TMD), cluster separately from other sub-families. Interestingly however, they diverge at different points within the phylogenetic tree, suggesting that these two TMD-lacking sub-families have arisen as independent loss-of-function (TMD) events.

We have also observed that those sequences that encode proteins associated with the MDR phenotype appear to segregate within the phylogeny. Within the B sub-family, a clear separation of ABCBs 1, 4, 5, and 11 can be seen from the rest of the sub-family (Figure 2), with the former group all being previously demonstrated to play a role in the development of MDR (Childs et al., 1998; Ambudkar et al., 1999; Smith et al., 2000; Huang et al., 2004). In addition, ABCG2 separates from the rest of the G sub-family, and is the only sub-family member associated with the MDR phenotype (Cervenak et al., 2006). Finally, the C sub-family also divides into those sequences associated with the MDR phenotype and those not, although in this case the segregation is not as clear. It is of interest to note that the ABCC9 and ABCC8 sequences appear to segregate with the MDR phenotype group, lending further weight to the suggestion that these genes may indeed contribute to a drug resistance phenotype (Deeley et al., 2006; Zhou et al., 2008).

To complement the nucleotide-level analysis, we also undertook a protein-level analysis for the human ABC transporters. In this analysis we also included the protein sequences from the pre-clinical species rat, mouse, dog and macaque. The derived phylogenetic tree is consistent with the orthologue assignment, with tight clustering of orthologues at the end of branches (Figure S6). In addition, simplification of the phylogenetic tree to illustrate the overall structure (Figure 3) is consistent with the conclusions drawn from the nucleotide-level analysis of human sequences. Whilst the multi-species amino acid-level analysis produces a different tree topology, clustering of MDR-associated sequences is still observed within the phylogeny (Figure 3).

DMD Fast Forward. Published on April 16, 2012 as DOI: 10.1124/dmd.112.045062
This article has not been copyedited and formatted. The final version may differ from this version.

Discussion

Proteins encoded by the A-D and G sub-families of the ABC transporter super-family play a central role in chemical ADME, affecting a compounds PK profile and, potentially its clinical efficacy (Hembruff et al., 2008; Kalliokoski and Niemi, 2009). It is important to understand the phylogenetic relationships of this super-family for two reasons: Firstly, the testing of novel chemical entities for both efficacy and toxicity is routinely undertaken in non-human mammalian species, with the data extrapolated to humans (Barille, 2008). In order for such extrapolations to be undertaken, complex physiologically-based pharmacokinetic models have been developed (Dressman et al., 2011). However, at present the role of drug transport proteins is poorly represented in many of these models, often being either encompassed in a generic 'active transport' term, or limited to very few specific transporters (Pang et al., 2009; Fan et al., 2010). One reason for this limitation is that the relationship between transporters in pre-clinical test species and humans is still relatively poorly understood, and as such this work aid in the focussing of experimental work to identify kinetic differences between orthologues. Once coupled with data on species differences in transporter expression (Takahashi et al., 2008; Cedernaes et al., 2011), this will allow far more robust cross-species extrapolation of drug ADME (Glavinas et al., 2004). in model species and humans. Secondly, understanding the mechanisms underlying MDR is an important step in identifying potential means to mitigate this important limitation to chemotherapeutic intervention (Coley, 2008).

Phylogenetic analysis over an extended evolutionary distance allows the clear assignment of sequences to sub-families for mammalian species. However, for species with a larger divergence time from humans, such as *Stronglocentrous purpuratus, Caenorhabditis elegans* and *Drosophila melanogastor*, robust sub-classification is not possible in the majority of

cases. Work by Sheps and colleagues also attempted to identify human orthologues for ABC drug transporters in *Caernorhabditis elegans,* using the amino acid sequence of ABC proteins (Sheps et al., 2004). They were able to assign orthologues to 8/43 human ABC drug transporters, and in the current study we confirm five of these using a different analysis methodology. The use of the entire coding sequence for phylogenetic analysis has several benefits over the use of only a selected region (e.g. TMD or ABD): Firstly, use of only a portion of the coding sequence excludes any variability seen within the rest of the sequence, which may result in bias in the generated phylogenies; secondly, whilst the TMD is most likely to be the most variable region, and hence main driver for the phylogenetic trees, use of this alone would exclude the non-TMD containing subfamilies (ABCE and ABCF), reducing the completeness of the analysis; thirdly, for robust assignment of orthologues through probabilistic orthology analysis it is crucial that all variability is accounted for within the analysis.

Probabilistic orthology analysis was able to robustly identify paralogues and orthologues between humans and several pre-clinical species, including the macaque which is currently poorly annotated. In addition, this analysis provides information on the rate of sequence change within the super-family, with gene duplication and loss rates of $0.0095Myr^{-1}$ and $0.0122Myr^{-1}$, respectively, being estimated for the ABC superfamily in humans, mouse, rat, dog and macaque. Cotton and Page have previously estimated that average duplication and loss rates in the vertebrate lineage over the last 200Myr to be $0.00115\ Myr^{-1}$ and $0.00749Myr^{-1}$ (Cotton and Page, 2005), meaning that for the ABC super-family both duplication and loss rates appear to be considerably higher than the average for all genes. It should be noted that other papers have estimated higher average duplication and loss rates (Lynch and Conery, 2000; Lynch and Conery, 2003), but there are potential confounders in these studies and, in

general, the averages are still lower than the estimate for the ABC super-family derived herein. These high duplication and loss rates supports the assignment of orthologue status via probabilistic analysis, as opposed to a simple MPR-based approach; Sennblad and Largergren demonstrated that the rate of false orthology predictions from an MPR-based approach increased with the duplication and loss rates (Sennblad and Lagergren, 2009). The presence of significantly higher duplications and loss rates for the ABC super-family could be reflective of a fluid phylogeny that can alter relatively rapidly, which would be logical for a protein family providing protection against chemicals in an ever-changing environmental milieu.

In comparison to previous publications, we have aligned the entire mRNA/protein sequence for phylogenetic analysis, rather than selected fragments. We demonstrate that robust phylogenies can be inferred from the alignment of full gene sequences (Supplemental Figures 1-5), distinguishing between full and half transporters within sub-families. In addition, for a single species, human, we have reconstructed the entire super-family. It is possible to successfully resolve the individual sub-families, and some interesting implications arise from this analysis. Aas noted within the introduction, two sub-families within the ABC super-family do not encode drug transporters, and indeed it has been argued that these genes should be excluded from the super-family (Rees et al., 2009). We demonstrate that these two sub-families have arisen by independent events, most probably through loss of the TMD. This loss of TMD has, obviously, led to an altered localisation of protein products from these sub-families, whilst their retention of an ABD allows them to undertake ATP-dependent processes. In the case of the ABCE sub-family, the sole gene encodes and ribonuclease L inhibitor, an important regulator of interferon action (Bisbal et al., 1995). In the case of the

three ABCF gene products, these proteins are all members of the GCN20 family and appear to play roles in TNFα-mediated signalling (Richard et al., 1998).

In addition, we provide data to support a clustering of those genes that encode transporters associated with MDR phenotype. This clustering, supported by both transcript and protein level analysis, could indicate that the MDR associated genes/proteins have features that set them apart from other genes within their sub-families. As this relationship translates to the protein level it also suggests that these features may be important in determining the molecular function(s) required to contribute to an MDR phenotype, although these features are as yet unelucidated. Whereas this separation from the main sub-family can be seen clearly in the B and G groups, it is undoubtedly less well defined within the C sub-family; this may be of interest considering that the B and G sub-family members encode proteins that generally have parent chemicals as their substrates, while those transporters encoded by the C sub-family generally transport conjugated products of metabolism (Choi, 2005; Deeley et al., 2006). Further examination is required to fully understand the impact of these different roles in chemical ADME, and on the development of a MDR phenotype.

In summary, the phylogenetic analyses contained herein extend current data on ABC gene orthologues in pre-clinical species, both identifying novel orthologues and correcting previous errors in annotation. Such information is important for the extrapolation of chemical effects in model organisms to humans and hence accurate risk assessment. In addition, we present the first complete human phylogeny across the entire super-family, demonstrating segregation between sequences that encode ABC transporters evoking the MDR phenotype and those which do not, at both the gene and protein level. This leads to the exciting possibility of focussing further on those transporters most likely to result in MDR and the development of strategies to mitigate this.

Author Contributions

Participated in Research Design: Plant, Fisher, Coleman

Conducted Experiments: Plant, Fisher

Performed Data Analysis: Plant, Fisher

Contributed New Reagents: Coleman

Wrote, or contributed to the writing of, the manuscript: Plant, Fisher, Coleman

References

Abascal F, Zaradoya R, and Posada D (2005) ProtTest: Selection of best fit models of protein evolution. *Bioinformatics* **21:**2104-2105.

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215:**403-410.

Ambudkar SV, Dey S, Hrycyna CA, Ramachandra M, Pastan I, and Gottesman MM (1999) Biochemical, cellular, and pharmacological aspects of the multidrug transporter. *Annual Review of Pharmacology and Toxiciology* **39:**361-398.

Arvestad L, Berglund AC, Lagergren J, and Sennblad B (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* **19:**i7-i15.

Barille FA (2008) *Principles of Toxicology Testing*. CRC Press, Boca Raton.

Bisbal C, Martinand C, Silhol M, Lebleu B, and Salehzada T (1995) Cloning and characterization of a rnase-l inhibitor - a new component of the interferon-regulated 2-5a pathway. *Journal of Biological Chemistry* **270:**13308-13317.

Cedernaes J, Olszewski PK, Almen MS, Stephansson O, Levine AS, Fredriksson R, Nylander O, and Schioth HB (2011) Comprehensive analysis of localization of 78 solute carrier genes throughout the subsections of the rat gastrointestinal tract. *Biochem Biophys Res Commun* **411:**702-707.

Cervenak J, Andrikovics H, Ozvegy-Laczka C, Tordai A, Nemet K, Varadi A, and Sarkadi B (2006) The role of the Human ABCG2 multidrug transporter and its variants in cancer therapy and toxicology. *Cancer Letters* **234:**62-72.

Childs S, Yeh RL, Hui D, and Ling V (1998) Taxol resistance mediated by transfection of the liver-specfic sister gene of P-glycoprotein. *Cancer Research* **58:**4160-4167.

Choi C-H (2005) ABC transporters as multidrug resistance mechanisms and the development of chemosensitizers for their reversal. *Cancer Cell International* **5:**30.

Coley HM (2008) Mechanisms and strategies to overcome chemotherapy resistance in metastatic breast cancer. *Cancer Treatment Reviews* **34:**378-390.

Cotton JA and Page RDM (2005) Rates and patterns of gene duplication and loss in the human genome. *Proceedings of the Royal Society B-Biological Sciences* **272:**277-283.

Davidson AL and Chen J (2004) ATP-binding cassette transporters in bacteria. *Annual Review of Biochemistry* **73:**241-268.

Dean M and Allikmets R (1995) Evolution of ATP-binding cassette transporter genes. *Current Opinion in Genetics & Development* **5:**779-785.

Dean M and Annilo T (2005) Evolution of the ATP-binding cassette (ABC) transporter superfamily in vertebrates. *Annual Review of Genomics and Human Genetics* **6:**123-142.

Deeley RG, Westlake C, and Cole SPC (2006) Transmembrane transport of endo- and xenobiotics by mammalian ATP-binding cassette multidrug resistance proteins. *Physiological Reviews* **86:**849-899.

Dressman JB, Thelen K, and Willmann S (2011) An update on computational oral absorption simulation. *Expert Opin Drug Metab Toxicol* **7:**1345-1364.

Fan JH, Chen S, Chow ECY, and Pang KS (2010) PBPK Modeling of Intestinal and Liver Enzymes and Transporters in Drug Absorption and Sequential Metabolism. *Curr Drug Metab* **11:**743-761.

Fath MJ and Kolter R (1993) ABC Transporters - Bacterial exporters. *Microbiological Reviews* **57:**995-1017.

Felenstein J (1989) Phylip-Phlylogeny Interface Package (Version 3.2). *Cladistics***:**164-166.

Felsenstein J (1997) An alternating least squares approach to inferring phylogenies from pairwise distances. *Systems Biology* **46:**101-111.

Gillet JP, Efferth T, and Remacle J (2007) Chemotherapy-induced resistance by ATP-binding cassette transporter genes. *Biochimica Et Biophysica Acta-Reviews on Cancer* **1775:**237-262.

Glavinas H, Krajcsi P, Cserepes J, and Sarkadi B (2004) The role of ABC transporters in drug resistance, metabolism and toxicity. *Current Drug Delivery* **1:**27-42.

Guindon S and Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52:**696-704.

Hembruff SL, Laberge ML, Villeneuve DJ, Guo BQ, Veitch Z, Cecchetto M, and Parissenti AM (2008) Role of drug transporters and drug accumulation in the temporal acquisition of drug resistance. *BMC Cancer* **8**.

Higgins DG and Sharp PM (1988) CLUSTAL - A package for performing multiple sequence aligment on a microcomputes. *Gene* **73:**237-244.

Huang Y, Anderle P, Bussey KJ, Barbcioru C, Shankavaram U, Dai Z, Reinhold WC, Papp A, Weinstein JN, and Sadee W (2004) Membrane transporters and channels: Role of the transportome in cancer chemosensitivity and chemoresistance. *Cancer Research* **64:**4294-4301.

Jacobs LL and Downs WR (1994) The evolution of murine rodents, in: *Rodents and Lagomorph Families of Asian Origin and Diversification* (Tomida Y, Li C, and Setoguchi T eds), pp 149–156., National Science Museum Monograph, Tokyo.

Kalliokoski A and Niemi M (2009) Impact of OATP transporters on pharmacokinetics. *British Journal of Pharmacology* **158:**693-705.

Koski LB and Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution* **52:**540-542.

Kumar S and Hedges SB (1998) A molecular timescale for vertebrate evolution. *Nature* **392:**917-920.

Le SQ and Gascuel O (2008) An improved general amino acid replacement matrix. *Molecular Biology and Evolution* **25:**1307-1320.

Lynch M and Conery J (2003) The evolutionary demography of duplicate genes. *Journal of Strutural and Functional Genomics* **3:**35-44.

Lynch M and Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290:**1151-1155.

Pang KS, Maeng HJ, and Fan JH (2009) Interplay of Transporters and Enzymes in Drug and Metabolite Processing. *Mol Pharm* **6:**1734-1755.

Rannala B and Yang ZH (2007) Inferring speciation times under an episodic molecular clock. *Systematic Biology* **56:**453-466.

Rees DC, Johnson E, and Lewinson O (2009) ABC transporters: the power to change. *Nat Rev Mol Cell Biol* **10:**218-227.

Richard M, Drouin R, and Beaulieu AD (1998) ABC50, a novel human ATP-Binding cassette protein found in tumor necrosis factor-alpha-stimulated synoviocytes. *Genomics* **53:**137-145.

Saier MH and Paulsen IT (2001) Phylogeny of multidrug transporters. *Seminars in Cell & Developmental Biology* **12:**205-213.

Sennblad B and Lagergren J (2009) Probabilistic Orthology Analysis. *Systematic Biology* **58:**411-424.

Sheps JA, Ralph S, Zhao ZY, Baillie DL, and Ling V (2004) The ABC transporter gene family of Caenorhabditis elegans has implications for the evolutionary dynamics of multidrug resistance in eukaryotes. *Genome Biology* **5**.

Smith AJ, van Helvoort A, van Meer G, Szabo K, Welker E, Szakacs G, Varadi A, Sarkadi B, and Borst P (2000) MDR3 P-glycoprotein, a phosphatidylcholine translocase,transports several cytotoxic drugs and directly interacts with drugs as judged by interference with nucleotide trapping. *Journal of Biological Chemistry* **275:**23530-23539.

Takahashi M, Washio T, Suzuki N, Igeta K, Fujii Y, Hayashi M, Shirasaka Y, and Yamashita S (2008) Characterization of gastrointestinal drug absorption in cynomolgus monkeys. *Mol Pharm* **5:**340-348.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, and Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25:**4876-4882.

Tiwari AK, Sodani K, Dai CL, Ashby CR, Chen ZS, and Chen ZS (2011) Revisiting the ABCs of Multidrug Resistance in Cancer Chemotherapy. *Curr Pharm Biotechnol* **12:**570-594.

Zhang Z, Schwartz S, Wagner L, and Miller W (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* **7:**203-214.

Zhou SF, Wang LL, Di YM, Xue CC, Duan W, Li CG, and Li Y (2008) Substrates and inhibitors of human multidrug resistance associated proteins and the implications in drug development. *Current Medicinal Chemistry* **15:**1981-2039.

Footnotes

This work was supported by AstraZeneca/UK Biotechnology and Biological Sciences Research Council [Grant BB/E527671/1].

DMD Fast Forward. Published on April 16, 2012 as DOI: 10.1124/dmd.112.045062
This article has not been copyedited and formatted. The final version may differ from this version.

Figure Legends

Figure 1: Gene duplication and loss likelihood scores within the ABC transporter encoding sub-families. Probabilistic Orthology analysis was undertaken for ABC-transporter proteins, using the gene evolution method of Arvestad (Arvestad et al., 2003). MCMC-estimated duplication and loss rates were calculated for every 10 iteration of a 10000 interation analysis.

Figure 2: A rooted consensus tree of the gene sequences for all human ATP-binding cassette super-family members. Phylogenetic tree was generated using the maximum-likelihood based program Phyml, with bootstrap values (100 replicates) shown at the major nodes. ABC genes whose proteins products have been positively associated with multiple drug resistance phenotype are highlighted in solid boxes.

Figure 3: A rooted consensus tree of the protein sequences sub-families A, B, C, D and G of the ATP-binding cassette super-family in humans, macaques, rat, mouse and dog. A multiple alignment was generated of ABC proteins from human, macaque, rat, mouse and dog using ClustalX2. Optimum amino acid replacement model was determined by ProTest, and then an LG algorithm and Phyml used to generate a phylogenetic tree using a maximum-likelihood approach. The full tree is presented as supplementary information (Figure S6), with a simplified cartoon showing only the overall general structure shown here. ABC genes whose proteins products have been positively associated with multiple drug resistance phenotype are highlighted in solid boxes.

DMD Fast Forward. Published on April 16, 2012 as DOI: 10.1124/dmd.112.045062
This article has not been copyedited and formatted. The final version may differ from this version.

Table 1: Total number of individual genes identified for each species.

	Sub-family								
Species name	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**?**	**Total**
Homo sapiens	12	11	12	4	1	3	5	-	48
Pan troglodytes	11	11	11	4	1	3	5	2	48
Macaca mulatta	13	11	12	4	1	3	3	-	47
Canis lupus familiaris	13	9	12	4	1	2	5	2	46
Mus musculus	14	11	11	4	2	2	5	2	51
Rattus norvegicus	14	10	9	4	1	2	5	3	48
Bos Taurus	11	9	10	3	1	3	5	3	45
Stronglocentrous purpuratus	1	4	6	1	1	2	2	12	29
Danio rerio	5	5	9	3	1	2	5	9	40
Caenorhabditis elegans	-	-	1	1	1	1	-	29	32
Drosophila melanogaster	-	-	-	2	1	1	1	38	43

Using human ABC transcript RefSeq sequences as the query term a cross-species mega-BLAST was undertaken to identify homologues in each of the listed species. Significant hits, not inclusive of transcript variants or pseudogenes, were included in an initial phylogenetic analysis, allowing designation of sequences into most probable subfamilies. Numbers in the '?' column indicate sequences identified in the megaBLAST search that cannot be conclusively identified as belonging to a specific sub-family.

Figure 1

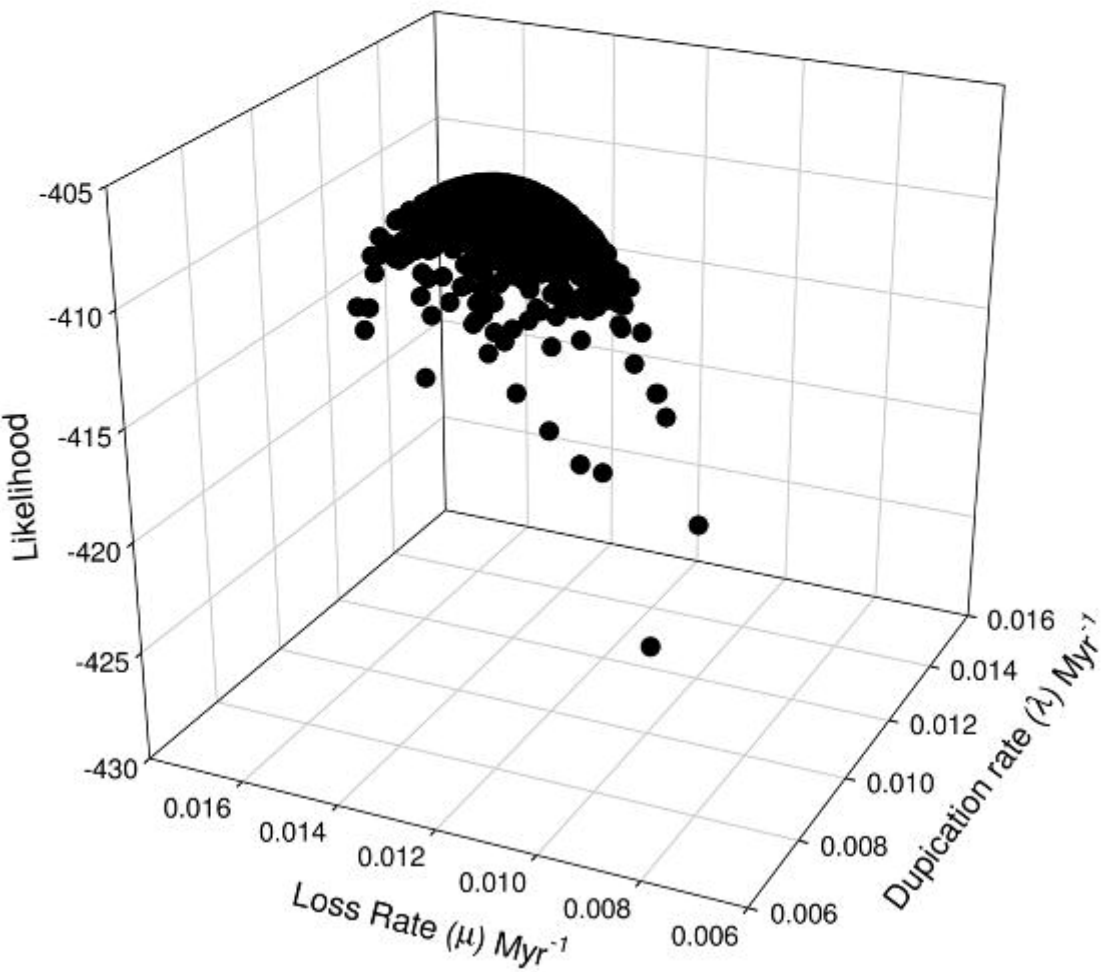

Likelihood
-405
-410
-415
-420
-425
-430
0.016
0.014
0.012
0.010
0.008
0.006
Loss Rate (μ) Myr-1
0.016
0.014
0.012
0.010
0.008
0.006
Dupication rate (λ) Myr-1

Figure 2

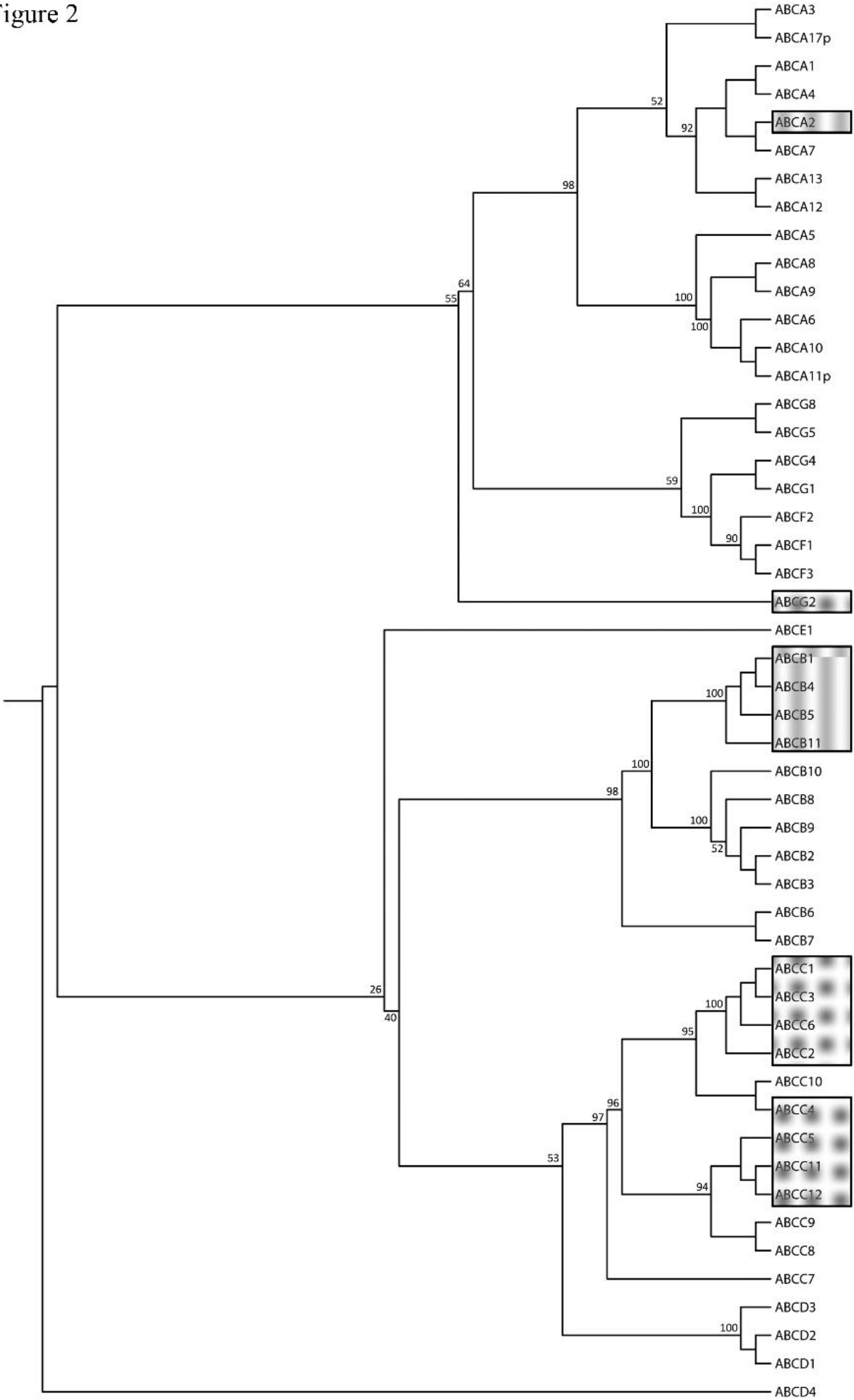
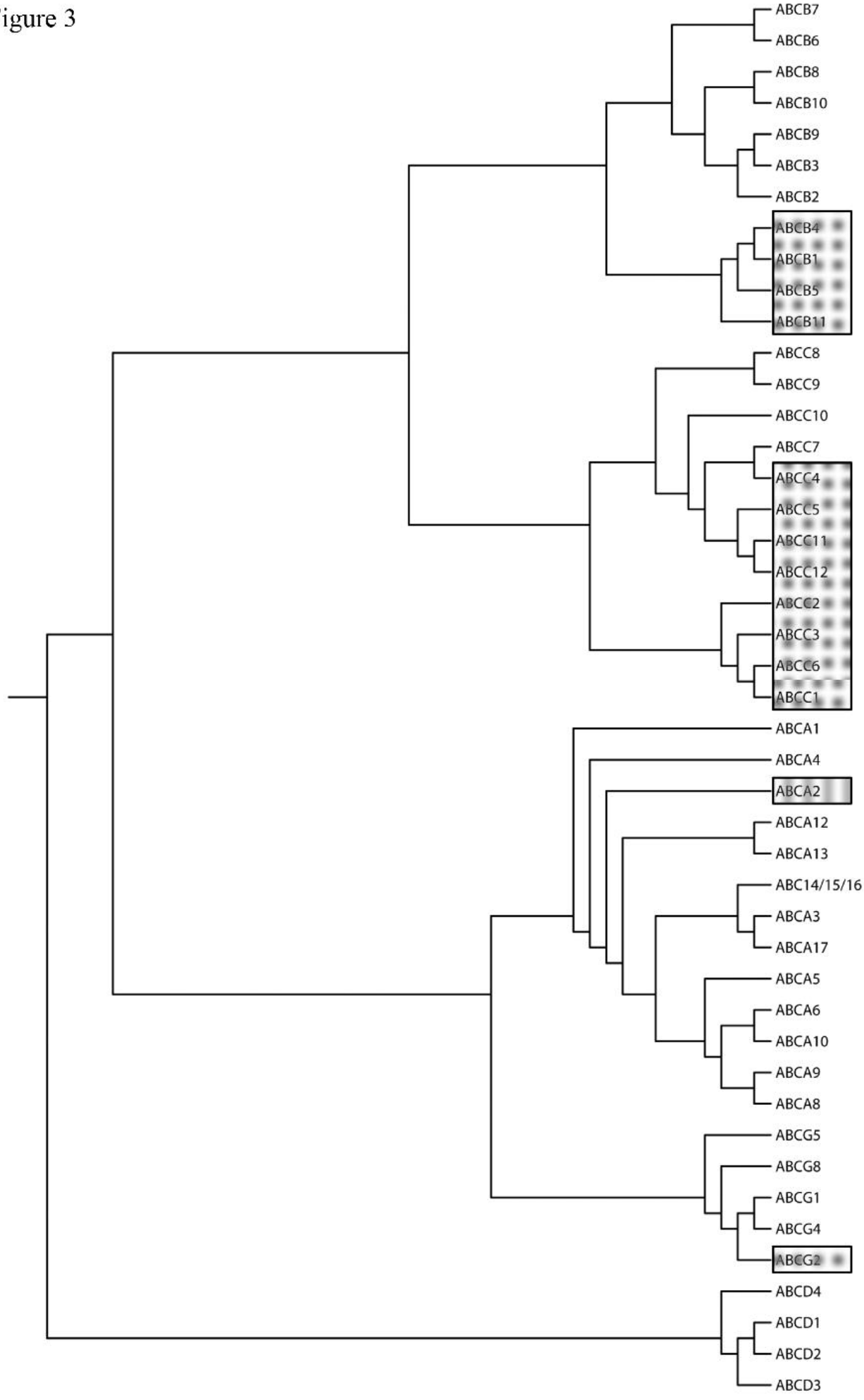

ABCA3
ABCA17p
ABCA1
ABCA4
ABCA2
ABCA7
ABCA13
ABCA12
ABCA5
ABCA8
ABCA9
ABCA6
ABCA10
ABCA11p
ABCG8
ABCG5
ABCG4
ABCG1
ABCF2
ABCF1
ABCF3
ABCG2
ABCE1
ABCB1
ABCB4
ABCB5
ABCB11
ABCB10
ABCB8
ABCB9
ABCB2
ABCB3
ABCB6
ABCB7
ABCC1
ABCC3
ABCC6
ABCC2
ABCC10
ABCC4
ABCC5
ABCC11
ABCC12
ABCC9
ABCC8
ABCC7
ABCD3
ABCD2
ABCD1
ABCD4
52
92
98
64
55
100
100
59
100
90
100
100
98
100
52
26
40
100
95
96
97
53
94
100

Figure 3

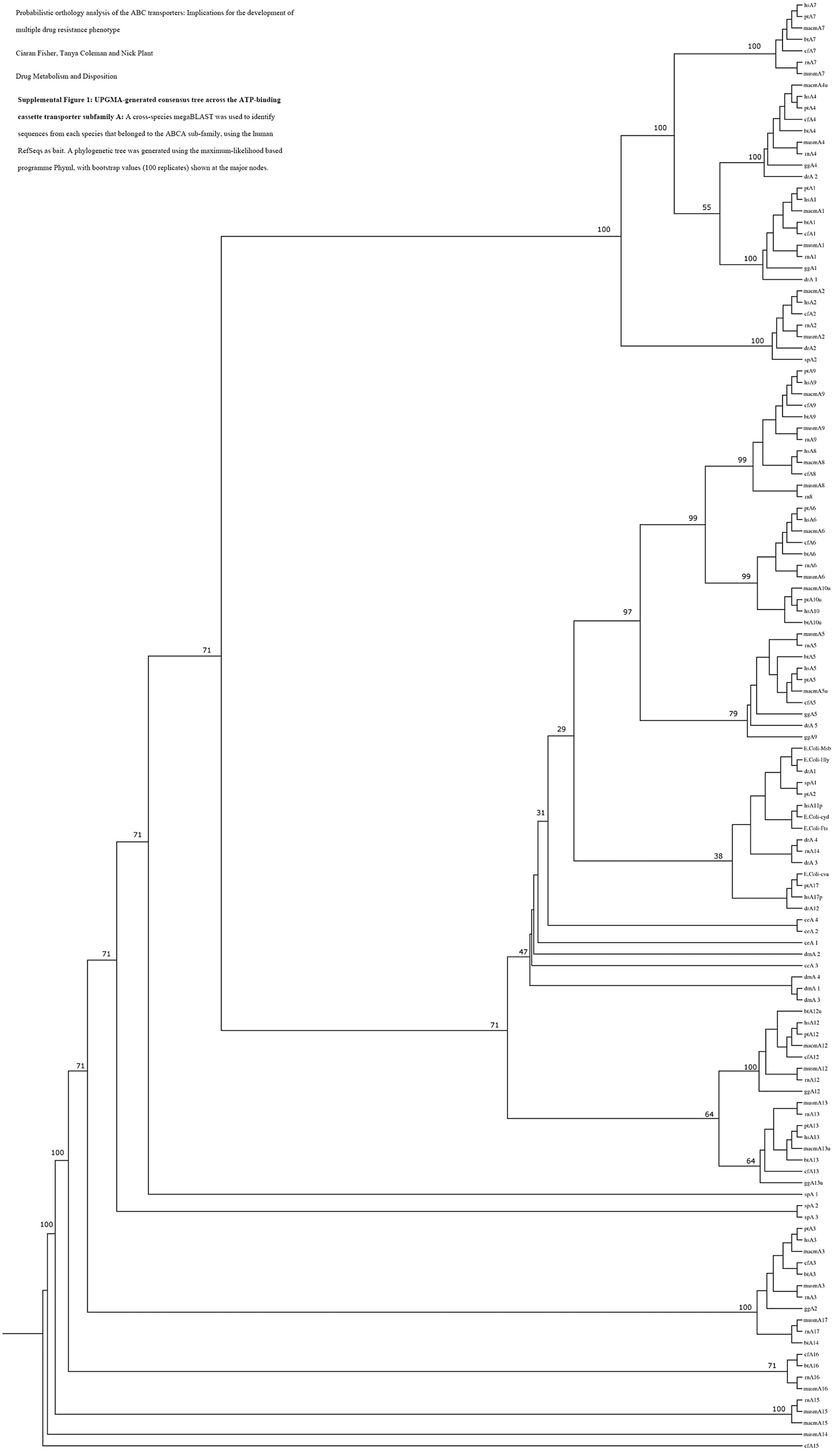

ABCB7
ABCB6
ABCB8
ABCB10
ABCB9
ABCB3
ABCB2
ABCB4
ABCB1
ABCB5
ABCB11
ABCC8
ABCC9
ABCC10
ABCC7
ABCC4
ABCC5
ABCC11
ABCC12
ABCC2
ABCC3
ABCC6
ABCC1
ABCA1
ABCA4
ABCA2
ABCA12
ABCA13
ABC14/15/16
ABCA3
ABCA17
ABCA5
ABCA6
ABCA10
ABCA9
ABCA8
ABCG5
ABCG8
ABCG1
ABCG4
ABCG2
ABCD4
ABCD1
ABCD2
ABCD3

Probabilistic orthology analysis of the ABC transporters: Implications for the development of multiple drug resistance phenotype

Ciaran Fisher, Tanya Coleman and Nick Plant

Drug Metabolism and Disposition

Supplemental Figure 1: UPGMA-generated consensus tree across the ATP-binding cassette transporter subfamily A: A cross-species megaBLAST was used to identify sequences from each species that belonged to the ABCA sub-family, using the human RefSeqs as bait. A phylogenetic tree was generated using the maximum-likelihood based programme Phyml, with bootstrap values (100 replicates) shown at the major nodes.

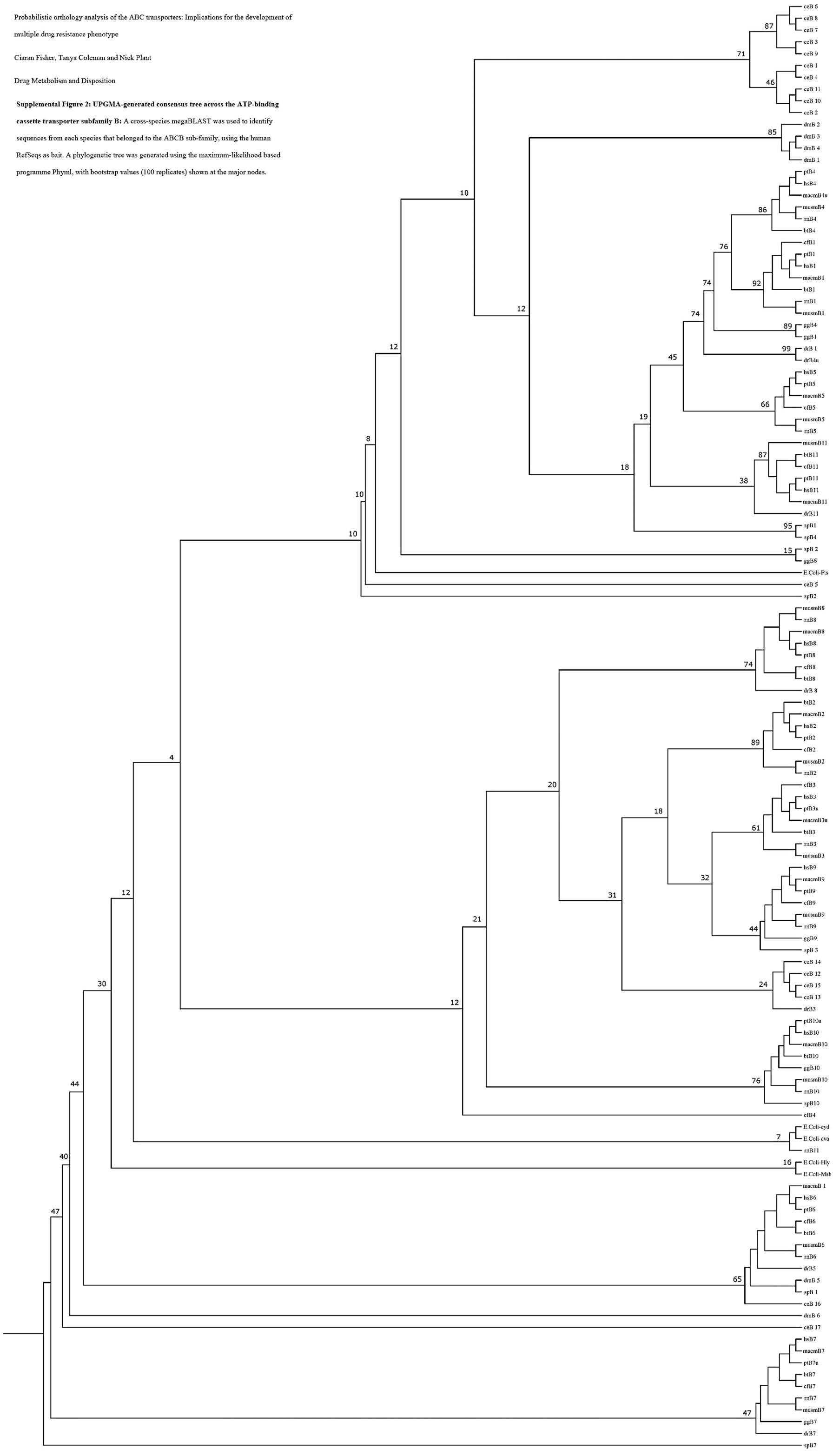


Probabilistic orthology analysis of the ABC transporters: Implications for the development of multiple drug resistance phenotype

Ciaran Fisher, Tanya Coleman and Nick Plant

Drug Metabolism and Disposition

Supplemental Figure 2: UPGMA-generated consensus tree across the ATP-binding cassette transporter subfamily B: A cross-species megaBLAST was used to identify sequences from each species that belonged to the ABCB sub-family, using the human RefSeqs as bait. A phylogenetic tree was generated using the maximum-likelihood based programme Phyml, with bootstrap values (100 replicates) shown at the major nodes.

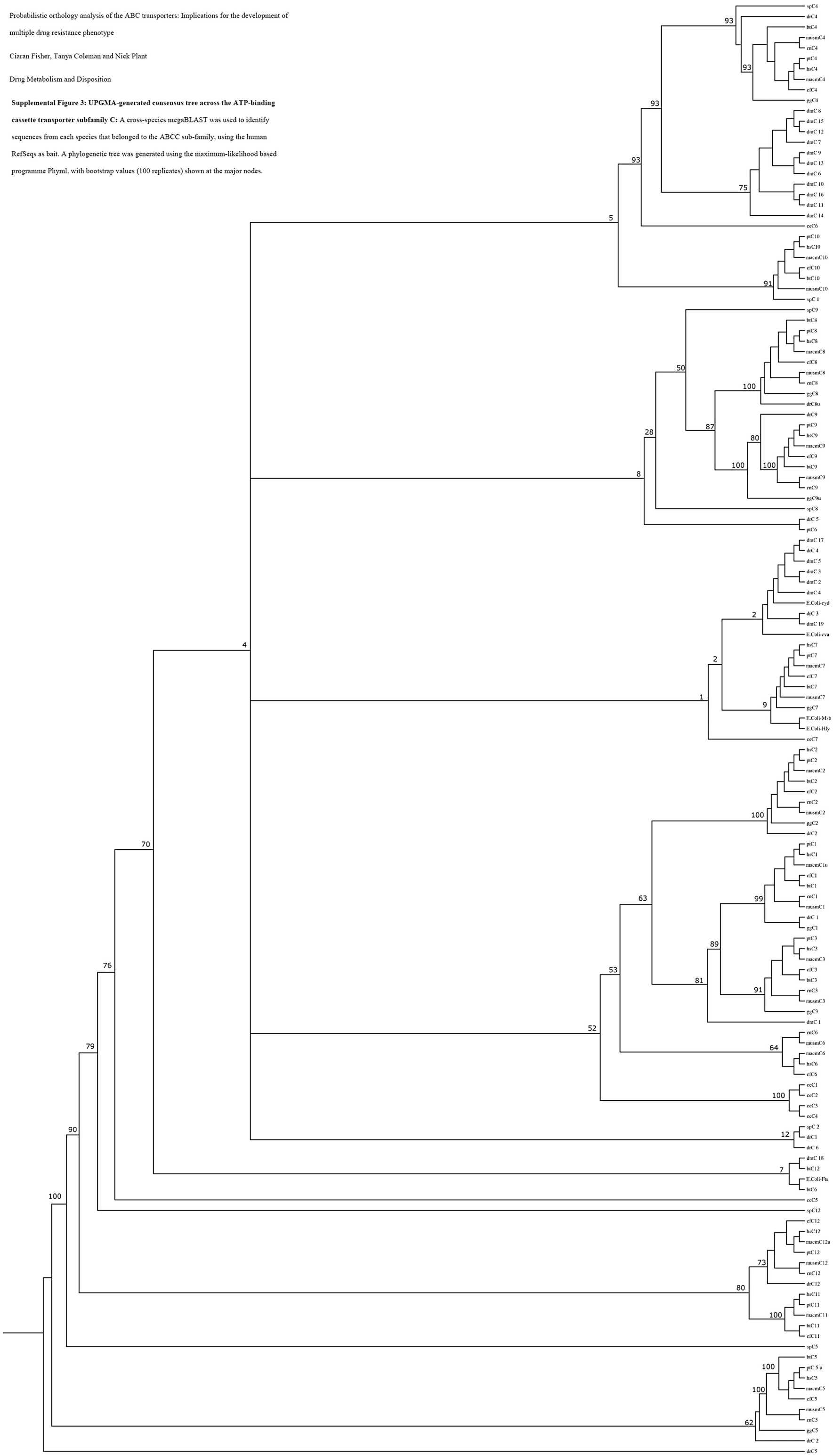


Probabilistic orthology analysis of the ABC transporters: Implications for the development of multiple drug resistance phenotype

Ciaran Fisher, Tanya Coleman and Nick Plant

Drug Metabolism and Disposition

Supplemental Figure 3: UPGMA-generated consensus tree across the ATP-binding cassette transporter subfamily C: A cross-species megaBLAST was used to identify sequences from each species that belonged to the ABCC sub-family, using the human RefSeqs as bait. A phylogenetic tree was generated using the maximum-likelihood based programme Phyml, with bootstrap values (100 replicates) shown at the major nodes.

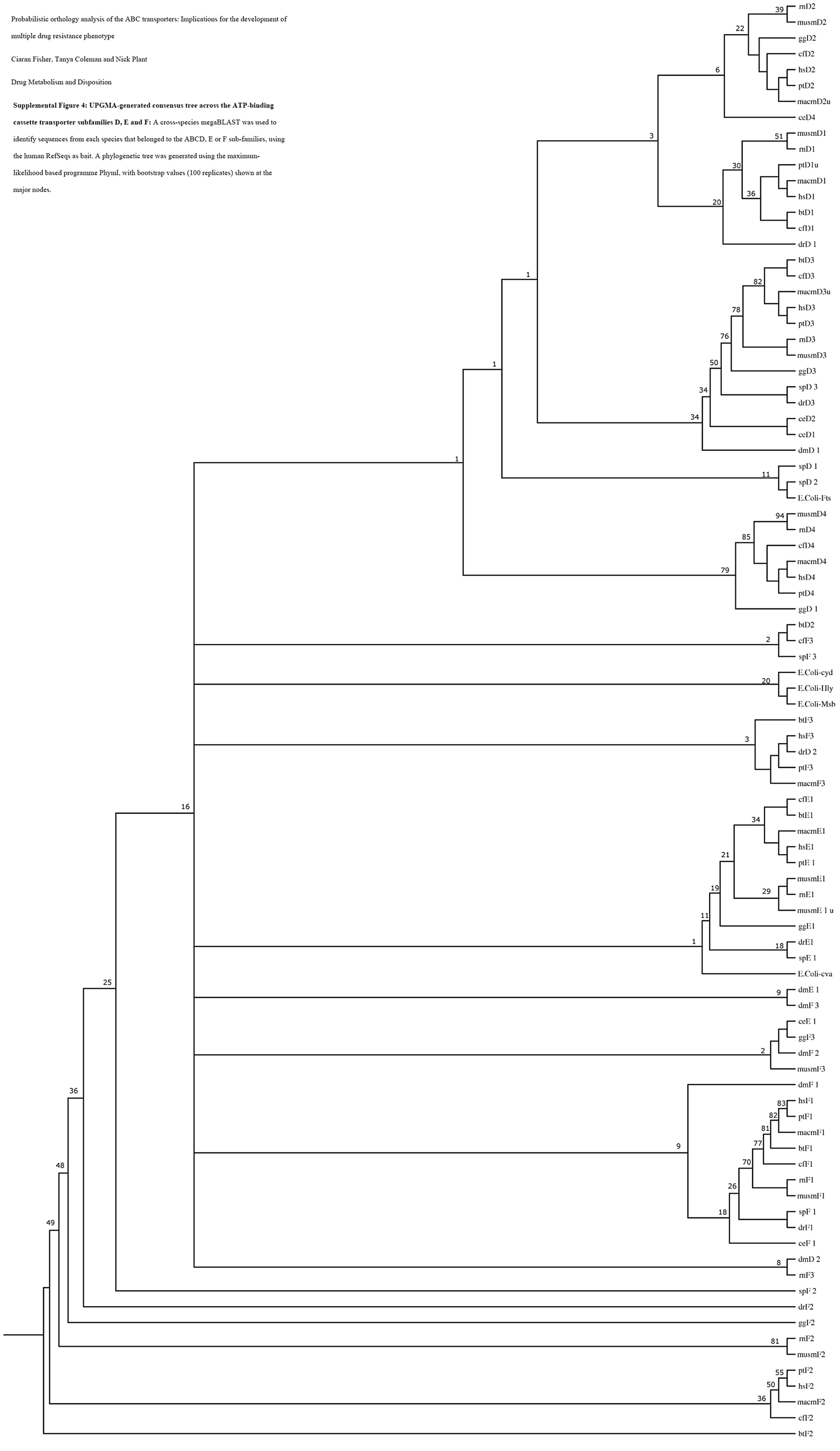


Probabilistic orthology analysis of the ABC transporters: Implications for the development of multiple drug resistance phenotype

Ciaran Fisher, Tanya Coleman and Nick Plant

Drug Metabolism and Disposition

Supplemental Figure 4: UPGMA-generated consensus tree across the ATP-binding cassette transporter subfamilies D, E and F: A cross-species megaBLAST was used to identify sequences from each species that belonged to the ABCD, E or F sub-families, using the human RefSeqs as bait. A phylogenetic tree was generated using the maximum-likelihood based programme Phyml, with bootstrap values (100 replicates) shown at the major nodes.

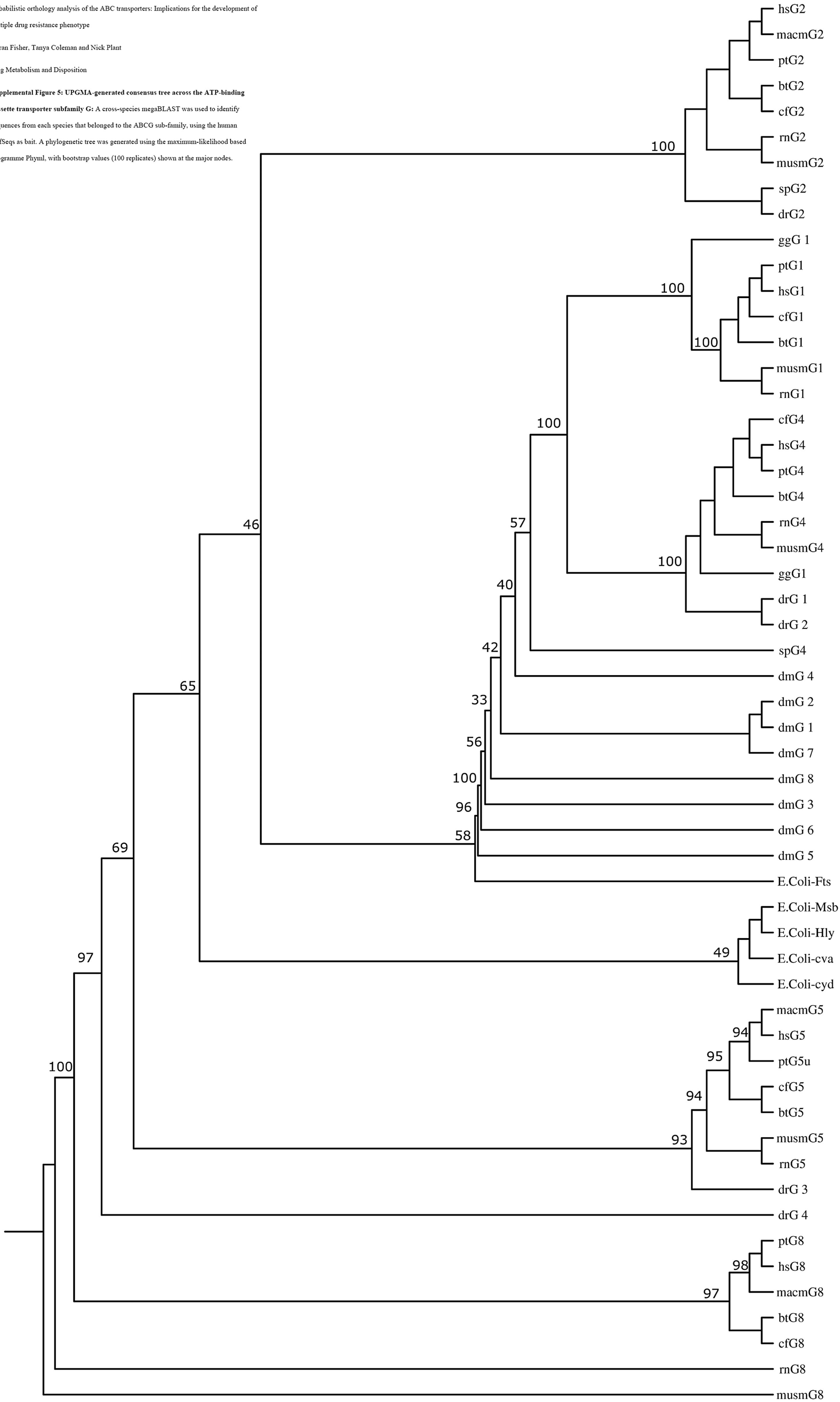


Probabilistic orthology analysis of the ABC transporters: Implications for the development of multiple drug resistance phenotype

Ciaran Fisher, Tanya Coleman and Nick Plant

Drug Metabolism and Disposition

Supplemental Figure 5: UPGMA-generated consensus tree across the ATP-binding cassette transporter subfamily G: A cross-species megaBLAST was used to identify sequences from each species that belonged to the ABCG sub-family, using the human RefSeqs as bait. A phylogenetic tree was generated using the maximum-likelihood based programme Phyml, with bootstrap values (100 replicates) shown at the major nodes.

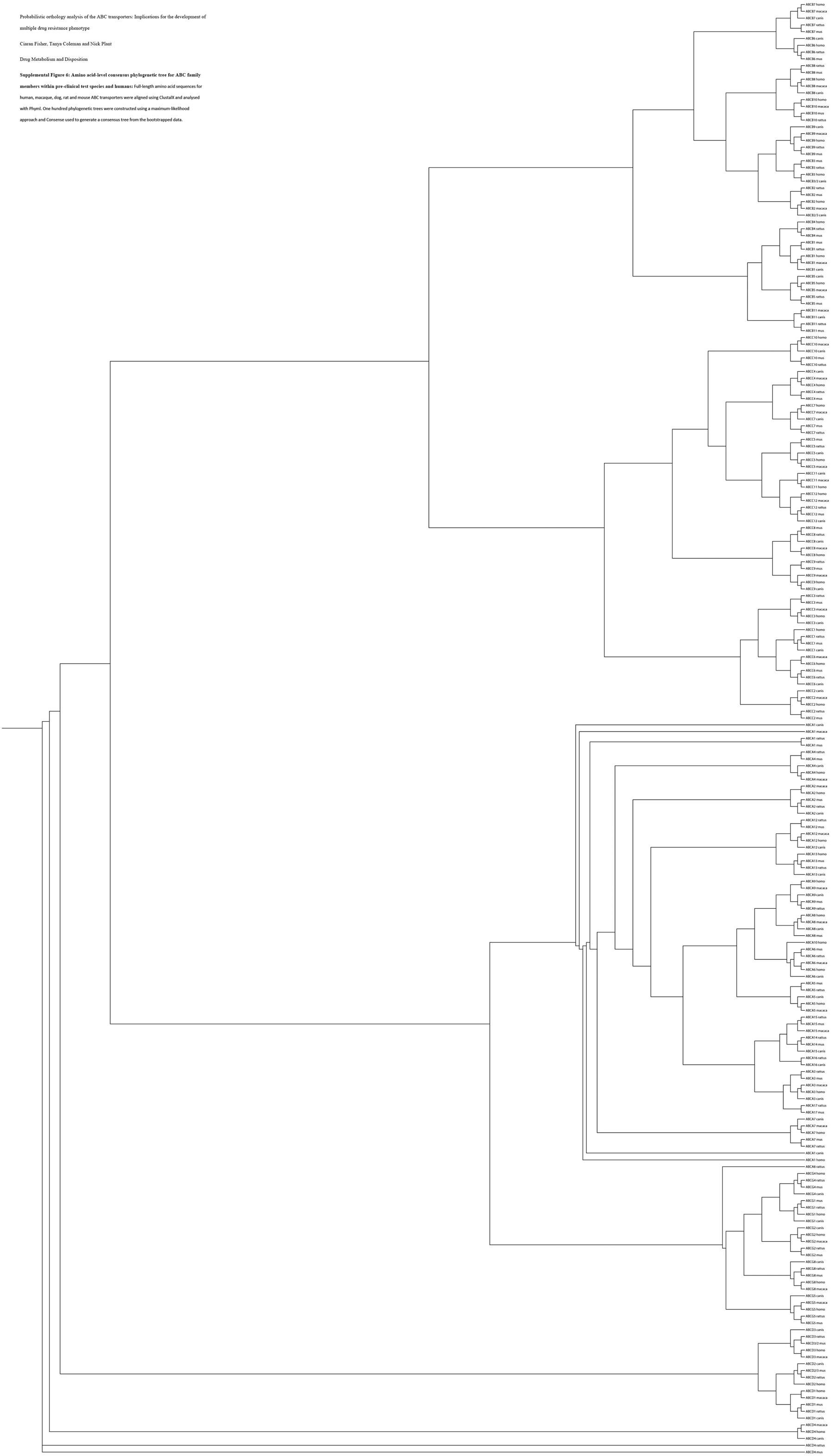


Probabilistic orthology analysis of the ABC transporters: Implications for the development of multiple drug resistance phenotype

Ciaran Fisher, Tanya Coleman and Nick Plant

Drug Metabolism and Disposition

Supplemental Figure 6: Amino acid-level consensus phylogenetic tree for ABC family members within pre-clinical test species and humans: Full-length amino acid sequences for human, macaque, dog, rat and mouse ABC transporters were aligned using ClustalX and analysed with Phyml. One hundred phylogenetic trees were constructed using a maximum-likelihood approach and Consense used to generate a consensus tree from the bootstrapped data.



Probabilistic orthology analysis of the ABC transporters: Implications for the development of multiple drug resistance phenotype

Ciaran Fisher, Tanya Coleman and Nick Plant

Drug Metabolism and Disposition

Supplemental Table 1: Human ABC super-family members.

All 48 members of the ABC super-family encoding genes found in humans, along with common alternate names, chromosomal location and RefSeq accession numbers (DNA and protein) used for an analysis. Note that pseudogenes, even if expressed, are not included in this table

Official Name	Alternate Name(s)	Chromosome Location	Genbank accession No.	
			Transcript	Protein
Sub-family A				
ABCA1	ABC1, FLJ14958	9q31.1	NM_005502	NP_005493
ABCA2	ABC2	9q34	NM_001606	NP_001597
ABCA3	ABC3	16p13.3	NM_001089	NP_001080
ABCA4	ABC10	1p22.1-21	NM_000350	NP_000341
ABCA5	ABC13, FLJ16381	17q24.3	NM_018672	NP_061142
ABCA6	FLJ43498	17q24.3	NM_080284	NP_525023
ABCA7	FLJ40025	19p13.3	NM_019112	NP_061985
ABCA8	MGC163152	17q24	NM_007168	NP_009099
ABCA9	MGC75415	17q24.2	NM_080283	NP_525022
ABCA10	EST698739	17q24	NM_080282	NP_525021
ABCA12	FLJ41584	2q34	NM_015657	NP_056472
ABCA13	FLJ16398	7p12.3	NM_152701	NP_689914

Sub-family B

ABCB1	P-glycoprotein (p-gp), MDR1	7q21.1	NM_000927	NP_000918
ABCB2	TAP1	6p21.3	NM_000593	NP_000584
ABCB3	TAP2	6p21.3	NM_000544	NP_000535
ABCB4	MDR2/3, p-gp3	7q21.1	NM_000443	NP_000434
ABCB5	EST422562	7p15.3	NM_178559	NP_848654
ABCB6	PRP, ABC14	2q36	NM_005689	NP_005680
ABCB7	ABC7, ASAT	Xq12-q13	NM_004299	NP_004290
ABCB8	MABC1	7q36	NM_007188	NP_009119
ABCB9	TAPL	12q24	NM_203444	NP_062570
ABCB10	M-ABC2	1q42.13	NM_012089	NP_036221
ABCB11	sister of p-glycoprotein (SPGP), bile salt transporter (BSEP)	2q24	NM_003742	NP_003733

Sub-family C

ABCC1	MRP1	16p13.1	NM_004996	NP_004987
ABCC2	MRP2	10q24	NM_000392	NP_000383
ABCC3	MRP3	17q22	NM_003786	NP_003777
ABCC4	MRP4	13q32	NM_005845	NP_005836
ABCC5	MRP5	3q27	NM_005688	NP_005679
ABCC6	MRP6	16p13.1	NM_001171	NP_001162
ABCC7	CFTR, MRP7	7q31.2	NM_000492	NP_000483
ABCC8	MRP8, SUR	11p15.1	NM_000352	NP_000343
ABCC9	SUR2	12p12.1	NM_005691	NP_005682
ABCC10	MRP7, SIMRP7	6p21.1	NM_033450	NP_258261
ABCC11	MRP8	16q12.1	NM_032583	NP_115972
ABCC12	MRP9	16q12.1	NM_033226	NP_150229

Sub-family D				
ABCD1	ALDP	Xq28	NM_000033	NP_000024
ABCD2	ALDRP	12q11-q12	NM_005164	NP_005155
ABCD3	PXMP1	1p22-21	NM_002858	NP_002849
ABCD4	PXMP1L	14q24.3	NM_005050	NP_005041
Sub-family E				
ABCE1	OABP	4q31	NM_002940	NP_002931
Sub-family F				
ABCF1	ABC27, ABC50	6p21.33	NM_001025091	NP_001020262
ABCF2	ABC28	7q36	NM_007189	NP_009120
ABCF3	FLJ11198	3q27.1	NM_018358	NP_060828
Sub-family G				
ABCG1	White1	21q22.3	NM_004915	NP_997513
ABCG2	Breast cancer resistance protein (BRCP)	4q22	NM_004827	NP_004818
ABCG4	White2	11q23.3	NM_022169	NP_071452
ABCG5	Sterolin 1, STSL	2p21	NM_022436	NP_071881
ABCG8	Sterolin 2, STSL	2p21	NM_022437	NP_071882

Probabilistic orthology analysis of the ABC transporters: Implications for the development of multiple drug resistance phenotype

Ciaran Fisher, Tanya Coleman and Nick Plant

Drug Metabolism and Disposition

Supplemental Table 2: Probabilistic Orthology Analysis

* Where orthologue predictions include two or more nodes within the phylogenetic tree, presented probabilities represent the product of the probabilities that each of these nodes represents a speciation event

†Where probabilistic orthology analysis indicates high probabilities for two human genes, the sequence is tentatively assigned as the orthologue to the human sequence with the highest probability, but alternate assignments are indicated by a forward slash.

ND = No potential orthologous sequence was identified from within the NCBI database

Human	Macaque	Mouse	Rat	Dog
Sub-family A				
ABCA1: NP_005493	XP_001106713 0.962	NP_038482 0.792*	NP_835196 0.792*	XP_538773 0.935
ABCA2: NP_001597	XP_001117819 0.982	NP_031405 0.945*	NP_077372 0.945*	XP_537788 0.959
ABCA3: NP_001080	XP_001085237 0.984	NP_038883 0.910*	XP_220219 0.910*	XP_537004 0.990
ABCA4: NP_000341	XP_002808277 0.984	NP_031404 0.913*	NP_001101191 0.913*	NP_001003360 0.977
ABCA5: NP_061142	XP_002800626 0.984	NP_671752 0.919*	NP_775429 0.919*	XP_537573 0.968*
ABCA6: NP_525023	XP_001083246 0.985	NP_671751 0.994	XP_001081607 0.994	XP_850922 0.986
ABCA7: NP_061985	XP_001093459 0.984	NP_038878 0.877*	NP_997481 0.877*	XP_542208 0.977
ABCA8: NP_009099	XP_001082492 0.984	NP_694785 0.769*	XP_221074 0.769*	XP_548020 0.990
ABCA9: NP_525022	XP_001082756 0.983	NP_671753 0.897*	XP_221101 0.897*	XP_853718 0.961
ABCA10: NP_525021	ND	ND	ND	ND
ABCA12: NP_056472	XP_001084970 0.984	NP_780419 0.862*	XP_001054709 0.862*	XP_536058 0.986
ABCA13: NP_689914	ND	NP_839990 0.994	NP_001099490 0.994	XP_848555 0.973
Sub-family B				
ABCB1: NP_000918	NP_001028059 0.984	NP_035205 0.852*	NP_036755 0.852*	NP_001003215 0.987
ABCB2: NP_000584	XP_001115506 0.984	NP_038711 0.919*	NP_114444 0.919*	XP_532099† 2/3:9.982
ABCB3: NP_000535	ND	NP_035660 0.954*	NP_114445 0.954*	† 3/2:0.988
ABCB4: NP_000434	ND	NP_032856 0.982	NP_036822 0.982	ND
ABCB5: NP_848654	XP_001102010 0.984	NP_084237 0.783*	XP_234725 0.783*	XP_539461 0.980
ABCB6: NP_005680	ND	NP_076221 0.993	NP_542149 0.993	XP_536073 0.965
ABCB7: NP_004290	XP_001097352 0.984	NP_033722 0.844*	NP_997683 0.844*	XP_549087 0.982
ABCB8: NP_009119	0.984	NP_083296 0.946*	NP_001007797 0.946*	XP_539916 0.984
ABCB9: NP_062570	XP_001096136 0.984	NP_063928 0.994	NP_071574 0.994	XP_849373 0.976
ABCB10: NP_036221	XP_001082734 0.985	NP_062425 0.980	NP_001012166 0.980	ND
ABCB11: NP_003733	XP_001097771 0.983	NP_066302 0.967*	NP_113948 0.967*	NP_001137404 0.983
Sub-family C				
ABCC1: NP_004987	ND	NP_032602 0.947*	NP_071617 0.947*	NP_001002971 0.981
ABCC2: NP_000383	NP_001028019 0.984	NP_038834 0.913*	NP_036965 0.913*	NP_001003081 0.975

Human	Macaque	Mouse	Rat	Dog
ABCC3: NP_003777	XP_001094709 0.984	NP_083876 0.895*	NP_542148 0.895*	XP_548204 0.980
ABCC4: NP_005836	XP_001085767 0.984	NP_001028508 0.878*	NP_596902 0.878*	XP_542642 0.983
ABCC5: NP_005679	0.984	NP_038818 0.968*	NP_446376 0.968*	NP_001121572 0.982
ABCC6: NP_001162	XP_001109862 0.983	NP_061265 0.876*	NP_112275 0.876*	XP_547113 0.946
ABCC7: NP_000483	0.984	NP_066388 0.833*	NP_113694 0.833*	NP_001007144 0.983
ABCC8: NP_000343	XP_001088694 0.982	NP_035640 0.914*	NP_037171 0.914*	XP_542520 0.981
ABCC9: NP_005682	XP_001098888 0.957	NP_066378 0.860*	NP_037172 0.860*	XP_543765 0.985
ABCC10: NP_258261	XP_001088553 0.984	NP_660122	NP_001101671	XP_538934 0.975
ABCC11: NP_115972	0.984	ND	ND	XP_535314 0.977
ABCC12: NP_150229	XP_002802515 0.987	NP_766500 0.943*	NP_955409 0.943*	XP_544420 0.987
Sub-family D				
ABCD1: NP_000024	XP_001085640 0.982	NP_031461 0.976*	NP_001102291 0.976*	XP_855341 0.976
ABCD2: NP_005155	ND	NP_036124† 2/3: 0.993	NP_036124 0.993	XP_534838 0.931
ABCD3: NP_002849	XP_002808274 0.985	NP_033017† 3/2:0.993	NP_036936 0.993	XP_537064 0.950
ABCD4: NP_005041	XP_001093730 0.984	NP_033018 0.654*	NP_001013118 0.654*	XP_547903 0.940
Sub-family G				
ABCG1: NP_058198	ND	NP_033723 0.994	NP_445954 0.994	XP_544902 0.987
ABCG2: NP_004818	NP_001028091 0.984	NP_036050 0.973*	NP_852046 0.973*	NP_001041486 0.987
ABCG4: NP_071452	ND	NP_620405 0.994	NP_001100286 0.994	XP_853231 0.987
ABCG5: NP_071881	XP_001111277 0.985	NP_114090 0.993	NP_446206 0.993	XP_538475 0.957
ABCG8: NP_071882	XP_001111321 0.984	NP_080456 0.993	NP_569098 0.993	XP_531799 0.966
Orthologues with no human version				
Sequence	**Species 1**	**Species 2**	**PO**	
ABCA14	Mouse: NP_080734	Rat: NP_001128063	0.990	
ABCA15	Mouse: NP_796187	Rat: NP_001099763	0.990	
ABCA15/14	Dog ABCA15: XP_547099	Rat ABCA14 NP_001128063	0.989	
ABCA16	Dog: XP_536943	Rat: XP_001079201	0.979	
ABCA17	Mouse: NP_001026792	Rat: NP_001026807	0.990	