GABRIEL DA ROCHA FERNANDES

INTEGRAÇÃO DE BASES DE DADOS DE GENES HOMÓLOGOS E APLICAÇÃO EM ANÁLISES DE SEQUÊNCIAS

BELO HORIZONTE
Março de 2011

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Bioinformática

INTEGRAÇÃO DE BASES DE DADOS DE GENES HOMÓLOGOS E APLICAÇÃO EM ANÁLISES DE SEQUÊNCIAS

Projeto de tese apresentado ao Curso de Pós-graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Bioinformática.

Orientador: Prof. Dr. José Miguel Ortega

Belo Horizonte
Março de 2011

Folha de aprovação

INTEGRAÇÃO DE BASES DE DADOS DE GENES HOMÓLOGOS E APLICAÇÃO EM ANÁLISES DE SEQUÊNCIAS

GABRIEL DA ROCHA FERNANDES

Tese defendida e aprovada pela banca examinadora constituída por:

Prof. Dr. José Miguel Ortega - Orientador
Universidade Federal de Minas Gerais

Prof. PhD. Peer Bork - Coorientador
European Molecular Biology Laboratory

Prof[a]. Dr[a]. Glória Regina Franco
Universidade Federal de Minas Gerais

Prof[a]. Dr[a]. Daniella Castanheira Bartholomeu
Universidade Federal de Minas Gerais

Prof. Dr. Sandro José de Souza
Ludwig Institute for Cancer Research

Prof Dr. Maurício Cantão
Laboratório Nacional de Computação Científica

AGRADECIMENTOS

Agradeço, em primeiro lugar, aos meus pais e irmão que deram amplo suporte e apoio durante todos os processos dessa vida acadêmica. Além, é claro, de toda a família.

Ao apoio, ensinamentos, e amizade do meu orientador José Miguel Ortega. Que me fez crescer muito com declarações como "A mente superior domina a inferior", ou "Isso e melhor do que escrever tese" seguido de uma foto de alguma festa.

Aos colegas do laboratório de biodados ao longos dos 6 anos e aos amigos de outros laboratórios pela convivência nos corredores e papos no café.

Aos amigos que compreenderam a ausência física nos momentos anti-sociais em que foram substituídos pelo trabalho.

Ao grupo do Dr. Peer Bork do EMBL-Heidelberg, na Alemanha, que me acolheu e ensinou muito ao longo do período em que estive lá.

"If I have seen further it is only by standing on the shoulders of giants."

Isaac Newton

RESUMO

Bases de dados biológicos são importantes fontes para pesquisas científicas. Algumas bases secundárias agrupam suas proteínas em grupos de ortólogos e categorias funcionais, como as bases COG (Cluster of Ortholog Groups) e KO (KEGG Orthology). A base KO foi usada em um teste de anotação automatizada de ESTs de *Caenorhabditis elegans*. Conduzimos um experimento controle em que a EST é designada à sua proteína cognata de *C. elegans*. Para a anotação simulamos um transcriptoma novo removendo as seqüências do verme da base de dados. Obtivemos três classes de anotação: corretas ou trocadas (quando o KO anotado era respectivamente igual ou discordante do designado) e especuladas (quando a EST era anotada, porém não designada). Obtivemos 68%, 4% e 28% de anotações corretas, trocadas e especuladas, respectivamente. Entretanto, as especulações diminuem para 4,4% quando designamos essas ESTs a proteínas que não estão na base KO. Para isso utilizamos proteínas KEGG não classificadas em grupos KO.

Na tentativa de aumentar a quantidade de informações em bases de dados como COG e KO, desenvolvemos uma metodologia baseada no recrutamento de seqüências que compartilhem o mesmo grupo UniRef50 de uma proteína recrutadora já existente na base de dados original. Um filtro de seleção de tamanho retirava recrutadas com mais que 10% de diferença de tamanho da recrutadora. Utilizando essa metodologia aumentamos a quantidade de proteínas na base COG de 124.369, provenientes de 63 genomas, para 961.725, com representantes de 3.477 genomas. A base recebeu a denominação *UniRef Enriched COG* (UECOG). Recentemente um novo enriquecimento foi feito utilizando um filtro em que exigíamos que o alinhamento entre a proteína recrutadora e a recrutada apresentasse valor-e menor que $1x10^{-10}$ e cobrisse pelo menos 50% da proteína recrutadora. Com isso obtivemos um total de 2.450.485 entradas, oriundas de 5.748 organismos distintos (UECOG 2.0).

O último procedimento foi utilizado para enriquecer a base de dados KO, aumentando as informações contidas de 1.940.617 proteínas para 4.447.538, e o número de organismos presentes de 1.315 para 32.213. A utilização de filtros de significância do alinhamento e de cobertura da seqüência recrutadora mostrou alta acurácia ao separar proteínas semelhantes, mas que possuem grupos de ortólogos distintos. A base enriquecida UEKO (*UniRef Enriched KO*) foi usada para testar a anotação automatizada de ESTs, como descrito anteriormente. A proporção de anotações trocadas diminuiu para 1% e as corretas aumentaram para 74%. Entretanto, as especulações continuaram freqüentes, mostrando que ainda existe muita informação a ser acrescentada. O número de anotações corretas, todavia, aumentou em 12%.

Foram realizados também estudos de metagenomas de microbiota intestinal humana. Um deles, utilizando 13 amostras públicas comparou as anotações proporcionadas pelo KO e UEKO. Essa comparação mostrou que a base UEKO anota mais que KO, já que mais de 100 grupos tem alinhamento exclusivo com a base enriquecida. Entretanto, a grande diferença é de caráter qualitativo, uma vez que há uma melhoria nos escores atribuídos pelo BLAST e as sequências são anotadas por proteínas de clados mais próximos, o que foi demonstrado por análise filogenética. O outro estudo procurou analisar, filogenética e funcionalmente, a estrutura da microbiota e identificamos nas amostras certos padrões filogenéticos e funcionais. Esses grupos, chamados de enterotipos, possuem características que os diferenciam dos demais, como a super-representação em um determinado enterotipo de enzimas envolvidas na síntese de vitaminas, em relação aos demais.

ABSTRACT

Biological databases are very useful sources for scientific research. Some secondary databases organize their data in orthologous groups and functional categories, such as COG (Cluster of Ortholog Groups) and KO (KEGG Orthology). The KO database was used for an automatic annotation test with *C. elegans*' ESTs. We performed a control experiment on which an EST is designated to its cognate protein in *C. elegans*. To the annotation stage we simulated a new transcriptome by removing the worm's sequences from the database. We obtained three annotation classes: correct or changed (when the annotated KO was equal or different from the designated, respectively) and speculated (when the EST is annotated, but not designated). We obtained 68%, 4% and 28% correct, changed and speculated annotations, respectively. However, the speculation decreases to 4,4% when we designate those EST using proteins that are not included in KO database.

Trying to increase the amount of information in databases like COG and KO, we developed a methodology based on recruiting sequences that share the UniRef50 cluster as a recruiter protein that is already present on the original database. A size selection filter removed recruited proteins with a difference higher than 10% the recruiter protein length. Using this methodology we increased the amount of proteins in the COG database from 124.369, from 63 genomes, to 961.725, representing 3.477 genomes. The database was denominated UniRef Enriched COG (UECOG). Recently a new enrichment was performed using a filter which we required that the alignment between the recruited and recruiter proteins showed an valor-e lower than $1x10^{-10}$ and cover at least 50% of the recruiter protein. We obtained 2.450.485 entries, from 5.748 distinct genomes (UECOG 2.0).

The previous procedure was used to enrich the KO database, increasing the amount of data from 1.940.617 proteins to 4.447.538, and the amount of organisms from 1.315 to 32.213. The usage of alignment significance filter and recruiter sequence coverage showed high accuracy in separating similar proteins, but with different orthologous groups. The enriched database UEKO (UniRef Enriched KO) was used to test the automated annotation of ESTs, as described previously. The proportion of changed annotation decreased to 1% and the correct increased to 74%. However, the speculation remained frequent, showing that we still have a lot of information to be added. The amount of correct annotation increased in 12%.

We also performed studies of the human gut microbial metagenome. One of them, using 13 public samples, compared the annotation provided by KO and UEKO. This comparison showed that the UEKO database annotates more sequences than KO, once that

more than 100 groups have exclusive alignment with the enriched database. However, the major difference is in qualitative aspect, once that we have an improvement in BLAST scores and proteins from closer clades annotate the sequences, which was demonstrated by phylogenetic analysis. The other study aimed in analyzing, phylogenetic and functionally, the microbiota structure and we identified certain phylogenetic and functional patterns. Those groups, known as enterotypes, have some features that differentiate them from the others, such as the over-representation of enzymes related to vitamin biosynthesis in some enterotype when compared to the others.

SIGLAS E ABREVIATURAS

BLAST	*Basic local alignment search tool*
CAMERA	*Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis*
CD-HIT	*Cluster Database at High Identity with Tolerance*
CDD	*Conserved Domain Database*
cDNA	DNA codificante
COG	*Cluster of Orthologous Groups*
DDBJ	*DNA Data Bank of Japan*
DNA	Ácido Desoxirribonucleico
dNTP	Desoxirribonucleotídeo fosfatado
EC	*Enzyme comission*
eggNOG	*evolutionary genealogy of genes: Non-supervised Orthologous Groups*
EMBL	*European Molecular Biology Laboratory*
EST	*Expressed sequence tag*
FTP	*File Transfer Protocol*
GAPDH	Gliceraldeído 3-fosfato desidrogenase
GI	*General Identifier*
HMM	Modelos ocultos de Markov
KAAS	*KEGG Automatic Annotation Server*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
KO	*KEGG Orthology*
KOALA	*KEGG Orthology and Links Annotation*
KOBAS	*KEGG Orthology Based Annotation System*
mRNA	RNA mensageiro
NCBI	*National Center for Biotechnology Information*
OrthoMCL	*Orthology Markov Cluster Algorithm*
PAM	*Partitioning Around Medoids*
PCA	Análise dos componentes principais
PCR	Reação em cadeia da polimerase
Pfam	*Protein family*
PHP	*Personal Hypertext Preprocessor*
PSI-BLAST	*Position-Specific Iterative BLAST*
RNA	Ácido ribonucleico
rRNA	RNA ribossômico
SBI	*System Biology Institute*
SMASH	*Simple Metagenomics Analysis SHell*
sRNA	*small RNA*
TGICL	*TIGR Gene Indices clustering tools*
UECOG	*UniRef Enriched COG*
UEKO	*UniRef Enriched KO*
UniProtKB	*UniProt Knowledgebase*
URL	Localizador padrão de recursos

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

1.1 A ERA GENÔMICA E PÓS GENÔMICA

Há cerca de duas décadas, em meados dos anos 90, deu-se o início de uma era genômica com a publicação do primeiro genoma completo (Fleischmann, Adams *et al.*, 1995). Essa tendência continuou desenvolvendo-se e chegou ao sequenciamento do genoma humano (Lander, Linton *et al.*, 2001; Venter, Adams *et al.*, 2001). Até então todos esses dados eram sequenciados por alguma versão do método de Sanger (Sanger, Air *et al.*, 1977; Sanger, 1981), a maioria por um processo semi-automático com eletroforese capilar, como pode ser visto na Figura 1 (A).

Hoje, em março de 2011, o NCBI conta com dados de 1001 organismos com genomas completos, 921 em fase de montagem, e 940 projetos ainda em andamento. Esse avanço deve-se muito ao desenvolvimento de estratégias alternativas para o sequenciamento de DNA, como o sequenciamento de arranjos cíclicos (Margulies, Egholm *et al.*, 2005; Shendure, Porreca *et al.*, 2005)

Essas novas alternativas levaram ao surgimento da "segunda geração" de sequenciadores, que se baseiam em melhorias no sistema de sequenciamento de arranjos cíclicos. Hoje existem diversas variações comerciais desse sistema que basicamente consiste em sequenciar um denso arranjo de DNA por ciclos interativos de manipulação enzimática e os dados são coletados por meio de imagens (Mitra e Church, 1999; Mitra, Shendure *et al.*, 2003), como pode ser visto na Figura 1 (B).

Vantagens da "segunda geração" estão na diminuição expressiva dos custos do sequenciamento, como mostrado na Tabela 1. Porém tem como desvantagem o pequeno tamanho das sequências geradas, criando aí um desafio para os bioinformatas.

Tabela 1 - Custos e tamanhos das sequências geradas por diferentes plataformas de sequenciamento.

Plataforma	Custo por Megabases (Dolares)	Tamanho das sequências (bases)
Sanger	~500	~800
454	~60	~350
Solexa	~2	~100
SOLID	~2	~75

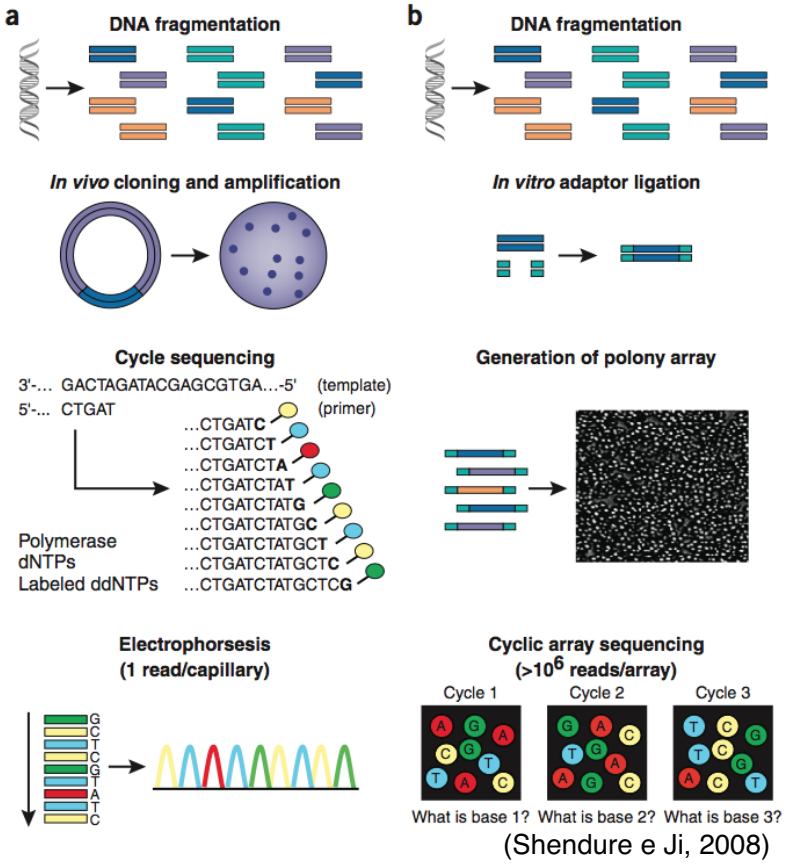

Figura 1 - Comparação dos processos de sequenciamento pelo método de Sanger e de segunda geração. (A) Etapas do sequenciamento pelo método Sanger, necessitando de clonagem in vivo, amplificação utilizando dNTPs e uma eletroforese capilar para a leitura das bases. Em (B) temos os métodos de segunda geração que dispensam a clonagem in vivo, a PCR ocorre em colônias no arranjo e a cada ciclo é tirada uma foto do sistema para a definição das bases.

Diante de grande quantidade de dados gerados começamos uma era pós genômica, em que a bioinformática tem papel fundamental em decifrar e organizar esses dados e associá-los com a biologia tradicional. Métodos foram desenvolvidos para análise de um grande número de genes e proteínas simultaneamente.

O primeiro passo para essa análise em larga escala foi o desenvolvimento do programa BLAST (Altschul, Gish *et al.*, 1990). Junto com o aumento da informação a ser analisada foram sendo desenvolvidas novas ferramentas como PSI-BLAST (Altschul, Madden *et al.*, 1997), aplicações de modelos ocultos de Markov (HMM) (Krogh, Brown *et al.*, 1994) e programas de alinhamentos múltiplos Clustal (Larkin, Blackshields *et al.*, 2007). Genes identificados por similaridade formavam uma longa lista e eram escolhidos genes de interesse para análises funcionais.

Entretanto, o aumento da quantidade de dados não leva necessariamente ao aumento do conhecimento biológico, a menos que este esteja acompanhado da melhoria das

ferramentas disponíveis (Bork e Koonin, 1998). E com o aumento de sequências de genomas completos, as bases de ortólogos começam a se tornar indispensáveis para a anotação funcional (Tatusov, Koonin *et al.*, 1997). Hoje, um importante papel da bioinformática, é o desenvolvimento de bases de dados secundárias contendo o conhecimento biológico acumulado até então para que possamos entender funções em níveis molecular, celular e sistêmico (Kanehisa e Bork, 2003).

1.2 TRANSCRIPTÔMICA

Transcriptoma é o conjunto de todos os trancritos de uma célula. Esses dados são analisados e comparadas em diferentes estágios de desenvolvimento, condições de crescimento, tecidos, e outros critérios. Entender o transcriptoma é essencial para a interpretação de elementos funcionais do genoma. O objetivo é catalogar todos os tipos de transcritos: mRNA, RNA não codificador e sRNA, determinar a estrutura transcricional do gene e modificações pós transcricionais.

Dentre as técnicas de análise temos a hibridação que consiste em incubar cDNA marcados com micro-arranjos gênicos (Ball, Sherlock *et al.*, 2002).

Existem outras técnicas baseadas no sequenciamento direto do cDNA. O sequenciamento pelo método de Sanger de cDNA em bibliotecas de EST (Boguski, Tolstoshev *et al.*, 1994), a Análise Serial da Expressão do Gene (SAGE) (Velculescu, Zhang *et al.*, 1995), e RNA-Seq (Mortazavi, Williams *et al.*, 2008) são alguns desses métodos.

ESTs são produtos do sequenciamento de extremidades do cDNA, ou de suas regiões internas – protocolo denominado ORESTES - e por isso são marcadores de sequências que estão sendo transcritas, daí o nome *Expressed Sequence Tag* (Adams, Kelley *et al.*, 1991). Um dos processos de geração de ESTs está descrito na Figura 2.

Por representarem regiões codificadoras freqüentemente truncadas, ou as vezes conterem apenas regiões UTR, essas sequências demandam de estratégias para sua montagem e a análise se beneficia muito de bases de dados pré-organizadas.

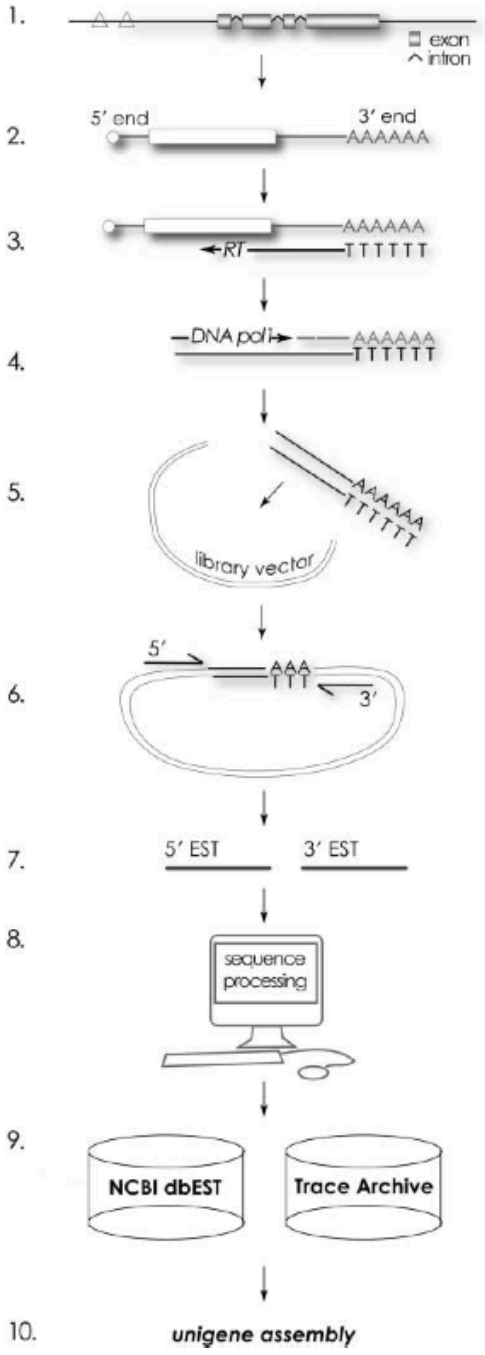

Figura 2 - Processo de produção e análise de seqüências EST. 1.Região de DNA genômico contendo introns e éxons e motivos reguladores (triângulos). 2. Introns são removidos do mRNA maduro no processo denominado edição; os mRNA são "encapados" na região 5' e caudas poli A são adicionadas na região 3'. 3. Transcriptase reversa é usada para a produção de DNA complementar (cDNA) pela molécula de mRNA. 4. Fita dupla de cDNA é produzida utilizando-se RNAse H e DNA polimerase. 5. Os cDNA são inseridos em vetores de clonagem para produzir uma biblioteca de cDNA. 6. Os insertos são seqüenciados por um ou ambos os lados (5' e/ou 3'). 7. As seqüências 5' e 3' resultantes são chamadas EST. 8. As EST são editadas para a remoção de seqüências de vetor, contaminantes e bases de baixa qualidade. 9. Depósito em bancos de dados públicos de EST (dbEST) e de cromatogramas (Trace Archive). 10. Passo alternativo de geração de unigenes**.** Fonte: (Bouck e Vision, 2007)**.**

1.3 ALINHAMENTO E ANOTAÇÃO DE SEQUÊNCIAS

A anotação de um genoma é geralmente citada como sendo o processo de identificação de genes (anotação estrutural) e predição de função a cada gene (anotação funcional), bem como todos os atributos biológicos que podem ser inferidos. A anotação funcional é geralmente executada utilizando bancos de dados de seqüências já conhecidas e programas de alinhamento entre seqüências como BLAST (Altschul, Gish *et al.*, 1990) ou algoritmos como Smith-Waterman (Smith e Waterman, 1981). O uso de um método adequado de alinhamento, assim como de bases de dados biológicas contendo informações como: ontologia, ortologia, vias metabólicas, e outros, são de suma importância para a qualidade da anotação (Thomas, Mi *et al.*, 2007).

1.3.1 Alinhamento local e global

Os dois principais métodos de comparar sequências são por meio de um alinhamento global e por alinhamento local. No global é feita uma tentativa de se alinhar toda a seqüência, enquanto no local são alinhadas subseqüências, ou regiões.

O uso de algoritmos de alinhamento global são a melhor forma de analisar relações evolutivas entre seqüências. Dentre eles, o mais conhecido é o de Needleman-Wunsch (Needleman e Wunsch, 1970). Hoje existem programas que fazem alinhamentos globais e múltiplos para facilitar essas análises evolutivas como o Clustal (Larkin, Blackshields *et al.*, 2007).

O alinhamento local é bastante usado para a anotação funcional, pois é capaz de identificar a existência de regiões conservadas (domínios conservados) entre seqüências. Os principais métodos para alinhamento local são o algoritmo de Smith-Waterman e o programa BLAST citados anteriormente.

O pacote de programas BLAST, distribuído pelo NCBI, é um hoje uma das ferramentas mais usados em bioinformática. Ele retorna alinhamentos locais, e por utilizar métodos heurísticos consegue ser mais rápido que o algoritmo de Smith-Waterman. Além disso, os programas BLAST são versáteis e permitem comparações entre sequências de nucleotídeos, entre aminoácidos, além do alinhamento de sequências de nucleotídeos traduzidas com sequências de aminoácidos. Uma importante característica do alinhador local BLAST é que, após ordenar os alinhamentos em ordem de escore, calcula o número de

alinhamentos de pontuação igual ou superior que poderiam ser obtidos sem nenhuma relação de ancestralidade, por mero acaso, e este número (valor-e) permite que o operador aceite ou não a hipótese alternativa ao simples acaso, ou seja, inferência de homologia.

1.4 BASES DE DADOS BIOLÓGICOS

Essa grande quantidade de dados gerada é armazenada em bases de dados biológicos. As principais bases de dados para depósitos de sequências são o GenBank do NCBI, EMBL e o DDBJ. Mas como foi dito anteriormente, a quantidade de dados não é diretamente proporcional à quantidade de conhecimento disponibilizado.

O NCBI é uma referência para bases de dados. Nesse centro encontramos bases para depósitos de sequências, consultas de artigos científicos, domínios conservados, e vários outras informações relevantes para a pesquisa biológica (Sayers, Barrett *et al.*, 2011).

Hoje existe uma enorme quantidade de bases de dados biológicos disponíveis. Cerca de 6% das bases de dados publicadas deixam de existir um ano após a publicação, e cerca de 20% das URLs do MEDLINE são inacessíveis (Wren e Bateman, 2008).

As relações de homologia entre os genes permitiram sua classificação em ortólogos e parálogos. Ortólogos são genes em espécies diferentes que evoluíram a partir de um ancestral comum por especiação, enquanto os parálogos são genes relacionados à duplicação de um gene dentro do mesmo genoma (Fitch, 1970). Uma vez que as funções tendem a ser conservadas em genes ortólogos, a informação de um gene pode então ser propagada para todo o grupo. Bases de dados que agrupam genes homólogos, ortólogos e parálogos, freqüentemente apresentam não somente os grupos de genes, mas categorias funcionais que compreendem esses grupos.

Assim, dentro desse universo de informações temos bases que se destacam em seguimentos específicos. Dentre as bases de dados que organizam as sequências em grupos de ortólogos temos o OrthoMCL-DB (Chen, Mackey *et al.*, 2006), eggNOG (Jensen, Julien *et al.*, 2008), KEGG Orthology (Kanehisa, Goto *et al.*, 2004) e COG (Tatusov, Koonin *et al.*, 1997). A organização e montagem de vias metabólicas pode ser feita com auxílio das bases KEGG Pathway (Kanehisa, 1997) e MetaCyc (Caspi, Foerster *et al.*, 2008). Termos de ontologia atribuídos a proteínas geram a base GOA (*Gene Ontology Association*) (Barrell, Dimmer *et al.*, 2009). Famílias e domínios protéicos são de extrema importância para a pesquisa biológica e para isso existem bases especializadas como Pfam (Bateman, Birney *et*

al., 2002), CDD (*Conserved Domain Database*) (Marchler-Bauer, Anderson *et al.*, 2005) e InterPRO (Hunter, Apweiler *et al.*, 2009).

1.4.1 KEGG

A base KEGG foi idealizada em 1995 (Kanehisa, 1997) para a organização do projeto genoma humano em vias metabólicas. Essa enciclopédia de genes e genomas é organizada em sub bases de dados, e no início contava apenas com as bases "Pathway", "Genes", "Enzyme" e "Compound".

O KEGG vem sendo desenvolvido e atualizado desde sua criação. Várias bases de dados foram criadas, como: "Genome" em 2000, "Reaction" em 2001, e em 2002 foi feita uma da mais relevantes melhorias que foi a criação de grupos de ortólogos em uma sub base chamada de KEGG Orthology (KO) (Kanehisa, Goto *et al.*, 2004).

A base de dados KO começou com o objetivo de integrar a informação genômica às redes protéicas através dos números EC, como identificadores comuns nos genomas e nas vias metabólicas. Hoje a informação é propagada utilizando análise computacional associada a uma curadoria manual (Kanehisa, Goto *et al.*, 2004). Os KO são peças importantes para a constituição de um sistema hierárquico do KEGG, mostrado na Figura 3. Em sua última versão (57.0 de 1 de Janeiro de 2011) essa hierarquia é composta por 7 categorias, 37 sub categorias, 388 vias e 14.323 grupos de ortólogos.

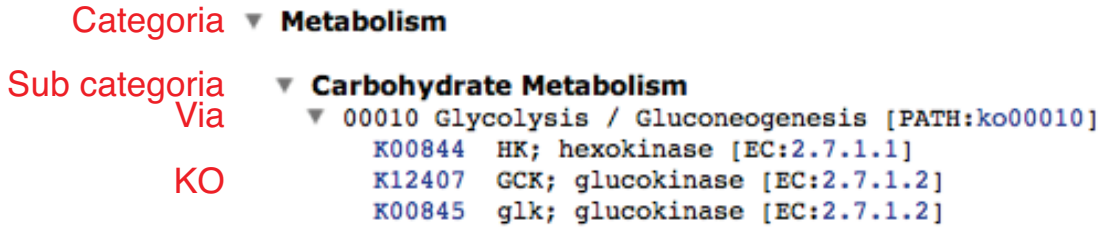

Figura 3 - Sistema de hierarquias funcionais do KEGG.

Recentemente o processo de anotação de ortólogos foi ampliado através do desenvolvimento da ferramenta KOALA (KEGG Orthology and Links Annotation). Existem dois métodos de anotação no KEGG. Um é baseado no genoma, em que curadores atribuem um grupo de ortólogo a genes de um dado genoma. O outro método é a propagação da anotação KO a genes de todos os organismos. Essas alternativas tiveram que ser adotadas diante da grande quantidade de dados que vem sendo adicionadas à base (Kanehisa, Goto *et al.*, 2010), evidenciando a necessidade de métodos automatizados de propagação da anotação já atribuída.

O KO é uma ferramenta importante para a anotação automatizada de sequências. Existem serviços como KAAS (Moriya, Itoh *et al.*, 2007) e KOBAS (Mao, Cai *et al.*, 2005) que fazem essa anotação automatizada baseada na estrutura dos grupos de ortólogos do KEGG.

1.4.2 COG e KOG

A base de dados COG (Cluster of Orthologous Groups) foi criada com o objetivo de extrair o máximo de informações dos genomas de organismos unicelulares que surgiam. Essa informação era obtida de acordo com as relações de homologia dos genes que compunham esses genomas. O entendimento de padrões na similaridade de sequências permitiu o agrupamento de proteínas nesses grupos de ortólogos. Cada grupo é constituído de proteínas ortólogas e algumas parálogas.

As relações de homologia entre os genes permitiu a classificação em ortólogos e parálogos. Ortólogos são genes em espécies diferentes que evoluíram a partir de um ancestral comum por especiação, enquanto os parálogos são genes relacionados à duplicação de um gene dentro do mesmo genoma (Fitch, 1970).

Uma vez que as funções tendem a ser conservadas em genes ortólogos, a informação de um gene pode então ser propagada para todo o grupo. E baseado nisso um método automático de agrupar genes em grupos de ortólogos contendo pelo menos 3 organismos foi criado (Tatusov, Koonin *et al.*, 1997).

A base de dados, que faz parte do NCBI, conta com informações de 66 genomas (63 deles procarióticos), contendo 192.987 proteínas, agrupadas em 4.872 COGs. Esses grupos são classificados em 25 categorias funcionais, que proporciona uma visão mais ampla dos processos desempenhados.

O mesmo processo de construção dos COGs foi adaptado para criar grupos para 7 genomas eucarióticos: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* e *Encephalitozoon cuniculi*. Um total de 60.759 proteínas desses organismos foram organizadas em 4.852 KOGs (os grupos de ortólogos de eucariotos). Genes que não estavam presentes em pelo menos 3 genomas eram organizadas em grupos especiais chamados de TWOGs (contendo 2 organismos) e LSE (contendo somente um organismo) (Tatusov, Fedorova *et al.*, 2003).

Logo após a publicação da base de dados KOG, ambas as bases não foram mais atualizadas. A última atualização da base COG aconteceu em setembro de 2003, por isso muitos de seus dados estão ultrapassados com entradas protéicas descontinuadas ou reanotadas. Mesmo assim, esse serviço do NCBI é um dos mais utilizados para anotação de genomas, principalmente microbianos.

1.5 UniProtKB e UniRef

Em 2002, os grupos do trEMBL, Swiss-Prot, EBI e PIR juntaram-se para formar o consórcio do UniProt. A missão primária era de dar suporte à pesquisa biológica fornecendo uma base de dados de alta qualidade, rica e com entradas precisamente anotadas e com várias referências a outras bases de dados (Apweiler, Bairoch *et al.*, 2004).

Os serviços prestados por esse consórcio divide-se em 3 principais bases de dados: UniParc, que agrupa uma coleção de sequências não redundantes por juntar toda a informação de sequências públicas; UniProtKB (UniProt Knowledgebase) que une informações das bases Swiss-Prot (anotação e curadoria manual) e trEMBL (anotação automatizada); e UniRef, que agrupas entradas que compartilham 50%, 90% ou 100% de identidade em uma sequência não redundante (Consortium, 2009).

Em sua mais recente versão, de março de 2011, o UniProtKB contava com 14.423.061 entradas protéicas, sendo 525.997 oriundas do Swiss-Prot e 13.897.064 do trEMBL. Essas proteínas estavam associadas a um total de 3.785.756 grupos UniRef50, com uma resolução de 50% de identidade entre as proteínas.

Os grupos UniRef50 são grupos que coletam entradas protéicas que compartilhem 50% de identidade entre eles e disponibilizam uma sequência referência e não redundante para cada grupo. Esses agrupamentos são formados pelo algoritmo CD-HIT (Li e Godzik, 2006) e são usados para limitar e agilizar as buscas em uma base tão grande que é o UniProtKB (Suzek, Huang *et al.*, 2007).

1.6 Integração de bases de dados

Diante dessa explosão de dados, torna-se cada vez mais necessária a propagação de informações a novas entradas em uma base de dados. Uma das formas dessa propagação é a integração de bases, que é dificultada com o aumento dos dados e falta de uma padronização na informação (Stein, 2003). Algumas informações são bem definidas e utilizadas em várias

bases de dados como o GI (General Identifier) e o Taxonomy ID atribuídos pelo NCBI. Eles são peças chave para os processos de integração.

Vários esforços são feitos por grandes grupos para a padronização da estrutura das bases de dados. Um projeto chamado Garuda é comandado pelo Dr. Hiroaki Kitano do SBI (System Biology Institute), e tem como objetivo a padronização e integração de dados para a construção de uma base para estudos de biologia sistêmica. O primeiro passo foi a criação de uma padronização para a representação dos dados de interações protéicas (Kitano, Funahashi *et al.*, 2005), o próximo passo é a aplicação dessa representação na montagem de vias com a propagação de informação para o maior número de entradas possível, com um processo de curadoria manual supervisionada pelo próprio usuário (Matsuoka, Ghosh *et al.*, 2010).

O desafio de integrar bases de dados de modo a automatizar, acelerar e facilitar a análise de dados é essencial para futuras análises em larga escala.

1.7 METAGENÔMICA

Cultivar uma bactéria em meios de cultura é geralmente o primeiro passo para o estudo desses microorganismos. Entretanto, as técnicas convencionais de cultivo são eficazes a menos que 1% da diversidade microbiana encontrada em amostras ambientais (Amann, Ludwig *et al.*, 1995).

O termo "metagenômica" descreve a análise coletiva, funcional e das sequências de micróbios existentes em amostras ambientais (Handelsman, Rondon *et al.*, 1998).

Existem duas vertentes de análise metagenômica. Uma baseia-se na expressão heteróloga dos genes clonados em uma biblioteca de DNA, faz-se então uma análise das funções celulares após a expressão. Outra vertente baseia-se em sequenciar todo o material genético amostrado e então analisá-lo.

Os primeiros estudos baseados em sequenciamento de DNA visavam apenas genes alvo como os codificadores do rRNA 16S e 18S. Isso permitia ter uma visão geral da composição e abundância filogenética do ambiente (Hugenholtz, Goebel *et al.*, 1998). Em 2003 o número de entradas no Genbank contendo sequências do gene de rRNA 16S de procariotos não cultiváveis era de 54.655, mais que o dobro do número dessas entradas oriundas de procariotos cultiváveis, que era 21.466 (Rappe e Giovannoni, 2003).

Embora ainda existissem muitos projetos visando o sequenciamento do rRNA 16S, um sequenciamento direto por "shotgun" poderia fornecer informações valiosas sobre o estilo de vida e capacidades metabólicas desses indivíduos não cultiváveis (Tringe, Von Mering *et al.*,

2005).

Em 2004, em um trabalho que marcou a metagenômica, Venter coletou amostras do mar do Sargasso e sequenciou utilizando a técnica de "shotgun". Nessas amostras ele encontrou pelo menos 1800 espécies diferentes, sendo 148 delas ainda desconhecidas. Além disso a análise funcional identificou 1,2 milhões de genes ainda desconhecidos e mais 782 novos fotorreceptores de rodopsina (Venter, Remington *et al.*, 2004).

A partir daí o campo foi evoluindo, e novas tecnologias foram sendo aplicadas, como o sequenciamento da microbiota intestinal de 124 indivíduos utilizando a plataforma Illumina. A análise desses dados permitiu a identificação de 3,3 milhões de genes diferentes existentes na flora intestinal (Qin, Li *et al.*, 2010).

Com o constante desenvolvimento e quantidade de dados gerados pela metagenômica surgem novos desafios para os bioinformatas. Diante de tanta informação ainda desconhecida, uma vez que 99,9% dos genomas da amostra não foram sequenciados, cabe aos pesquisadores desenvolverem técnicas para a montagem dos genomas, e desenvolvimento de bases de dados com uma cobertura satisfatória para a anotação do máximo de sequências possível (Raes, Foerstner *et al.*, 2007). Alguns *pipelines* para anotação metagenômica já estão disponíveis para auxiliar nesses desafios, dentre eles destacam-se o CAMERA (Seshadri, Kravitz *et al.*, 2007), MG-RAST (Meyer, Paarmann *et al.*, 2008) e SMASH (Arumugam, Harrington *et al.*, 2010), cujas etapas estão descritas na Figura 4.

O processo de caracterização de um ecossistema por meio da metagenômica tem várias etapas e elas estão interligadas, no sentido de que o resultado de uma etapa influencia no resultado de outra, e consequentemente na análise final (Raes e Bork, 2008). Essa rede de influências pode ser vista na Figura 5.

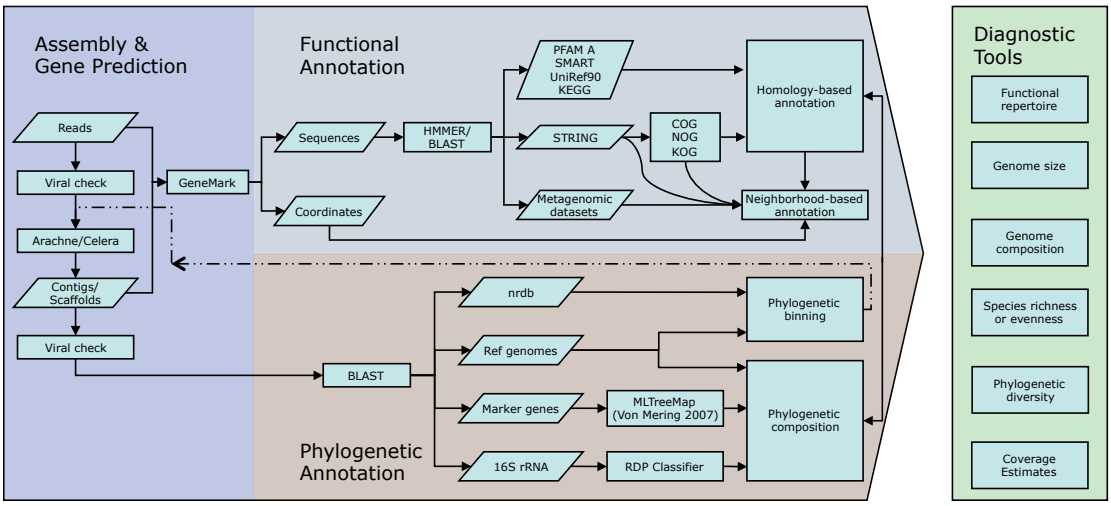

Figura 4 - Fluxo de análises do programa SMASH. O *pipeline* começa com as sequências iniciais e realiza a montagem e a predição dos genes. A anotação funcional utiliza os genes preditos em uma busca por homólogos em diversas bases de dados protéicas. A análise filogenética busca, utilizando BLAST, associações com sequências de bases públicas, bases locais de genomas referências, assim como genes marcadores e rRNA 16S.

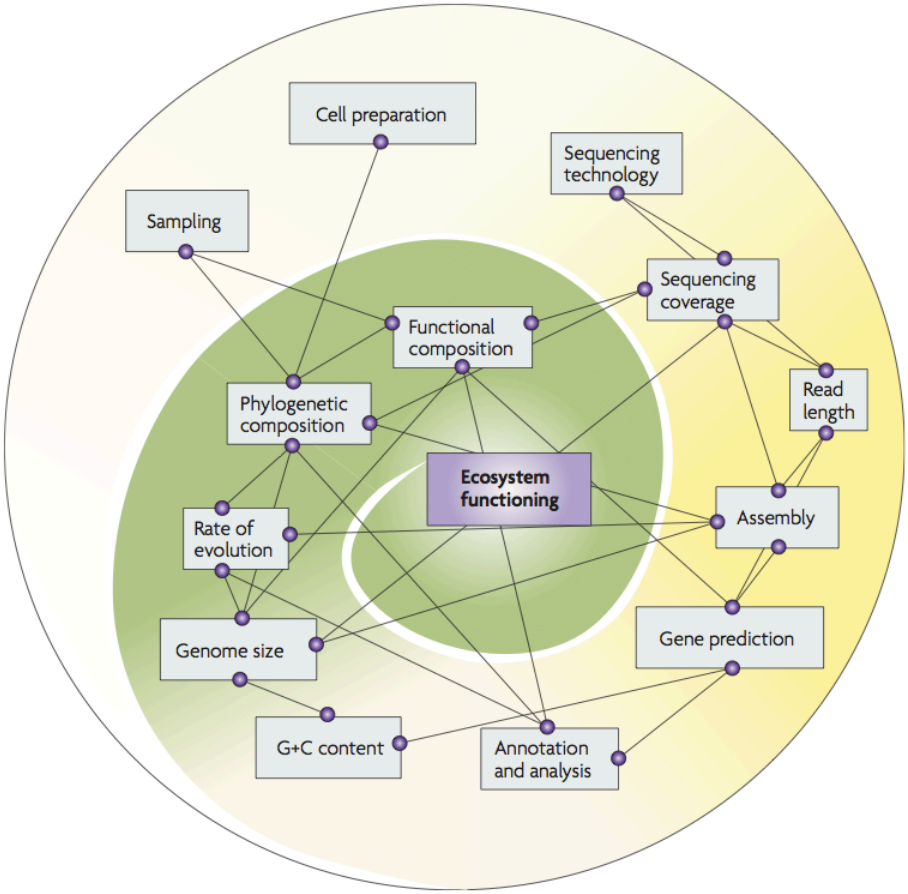

Figura 5 - Do metagenoma à funcionalidade do ecossistema: fatores influentes e dependências ocultas. Uma visão geral dos fatores necessários para análise do ecossistema. As linhas entres os fatores indicam interdependências, por exemplo, a amostragem é um fator chave pois influencia a análise de composição funcional e filogenética. Fonte: (Raes e Bork, 2008).

2 OBJETIVOS

- Avaliação das principais bases de dados de genes ortólogos.

- Desenvolvimento de uma metodologia para a integração dessas bases e propagação das suas informações.

- Utilização das bases criadas para análises de sequências de transcriptômica e metagenômica.

3 MÉTODOS

3.1 OBTENÇÃO DE SEQUÊNCIAS E INFORMAÇÕES

3.1.1 Criação de uma base KEGG local

As informações da base de dados KEGG foram obtidas do servidor FTP (ftp://ftp.genome.jp/pub/kegg). Informações que associem as proteínas aos seus GI e agrupamentos KO foram obtidas pelos arquivos "genes_ncbi-gi.list" e "genes_ko.list" respectivamente, ambos localizados nas pastas "/linkdb/genes/" desse mesmo servidor.

As sequências de proteínas em formato FASTA foram obtidas pelo arquivo "/fasta/genes.pep" do servidor FTP.

Também foram adquiridas informações sobre as vias em "/pathway/pathway" e sobre os módulos das vias em "/module/module". Essas informações consistem nos KOs que compões essas vias e módulos e suas definições.

Todos esses dados foram varridos para a extração de informações relevantes que foram armazenadas em um banco de dados MySQL em tabelas devidamente organizadas.

3.1.2 Criação de uma base COG local

Da mesma maneira que na criação da base KEGG local os dados foram obtidos pelo servidor FTP (ftp://ftp.ncbi.nih.gov/pub/COG/COG), para termos os GI associados às entradas foi utilizado o arquivo "myva=gb" e para associar aos grupos usamos o arquivo "whog" presentes nesse servidor. Mesmo local em que obtivemos as sequências no formato FASTA pelo arquivo "myva". Toda essa informação foi também armazenada em banco MySQL.

3.1.3 Criação de uma base UniProt local

Para a criação da versão local do UniProt usamos basicamente dois grupos de arquivos, encontrados no servidor FTP (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase). O primeiro deles

é um arquivo chamado "idmapping_selected.tab" e associa cada entrada a um grupo UniRef50, a um GI e a um identificador taxonômico do seu organismo. O outro são dois arquivos, um para a versão SwissProt e outro para a versão trEMBL do UniProt, chamados "uniprot_sprot.dat" e "uniprot_trembl.dat", respectivamente. Deles são retiradas informações sobre o número EC, evidências da existência da proteína, integridade e sequência de aminoácidos de cada entrada. Dados também armazenados no banco MySQL. A integridade das sequências é um aspecto importante a ser analisado, uma vez que existem entradas UniProt que dizem respeito a fragmentos de proteínas que são descartados das nossas análises.

3.1.4 Obtenção das ESTs

As ESTs utilizadas nos experimentos foram obtidas através de consulta ao NCBI, mas especificamente à base de dados EST, em que procuramos apenas pelo organismo *C. elegans* através da expressão "txid6239[Organism:noexp]" no campo de busca.

3.1.5 Amostras metagenômicas

As sequências públicas (Kurokawa, Itoh *et al.*, 2007) dos 13 indivíduos utilizados foram conseguidas através de colaboração com EMBL-Heidelberg e estão agora disponíveis no endereço http://www.bork.embl.de/Docu/human_gut_microbial_gene_catalog/

3.2 RECURSOS COMPUTACIONAIS E PROGRAMAS UTILIZADOS

A maior parte do trabalho analítico foi realizada em servidores com sistema operacional Linux, que nativamente disponibiliza o ambiente necessário para a realização do trabalho, com os programas PERL, MySQL, PHP.

Alguns outros foram necessários e então instalados, dentre eles o BLAST (Altschul, Gish *et al.*, 1990), Clustal (Larkin, Blackshields *et al.*, 2007), TGICL (Pertea, Huang *et al.*, 2003), GeneMark (Lukashin e Borodovsky, 1998), SeedLinkage (Barbosa-Silva, Satagopam *et al.*, 2008).

Cladogramas foram gerados pelo serviço iTOL disponível em http://itol.embl.de/ (Letunic e Bork, 2007) e os mapas metabólicos produzidos pela ferramenta iPath (Letunic, Yamada *et al.*, 2008) disponível em http://pathways.embl.de/iPath2.html.

3.3 TESTE DE ANOTAÇÃO

O teste de anotação deu-se seguindo a metodologia descrita no artigo publicado (Fernandes, Mudado *et al.*, 2008).

Para a primeira etapa, chamada de designação, que funciona como o experimento controle, utilizamos somente as proteínas de *C. elegans* das bases de dados a serem consultadas. Essas proteínas foram então utilizadas como base de dados para a consulta do BLAST, programa esse executado sob orientação de alguns parâmetros como a inativação do filtro de complexidade através da expressão "-F F", seleção somente dos resultados com valor-e menores que 10^{-10} e identidade entre as sequências de pelo menos 80% (Mudado e Ortega, 2006). Para cada EST aceitamos como resultado a proteína associada com melhor valor de escore.

Durante a próxima etapa, chamada de anotação, simulamos que o nossas ESTs pertencem a um novo genoma, para isso removemos as entradas de *C. elegans* da base que vai ser usada para o BLAST. A busca por homologia segue os mesmos parâmetros, exceto pela exigência de 80% de identidade, que agora não é requisito para a associação.

Os resultados são comparados e um fluxograma descrevendo o experimento e sua análise está disponível na Figura 6.

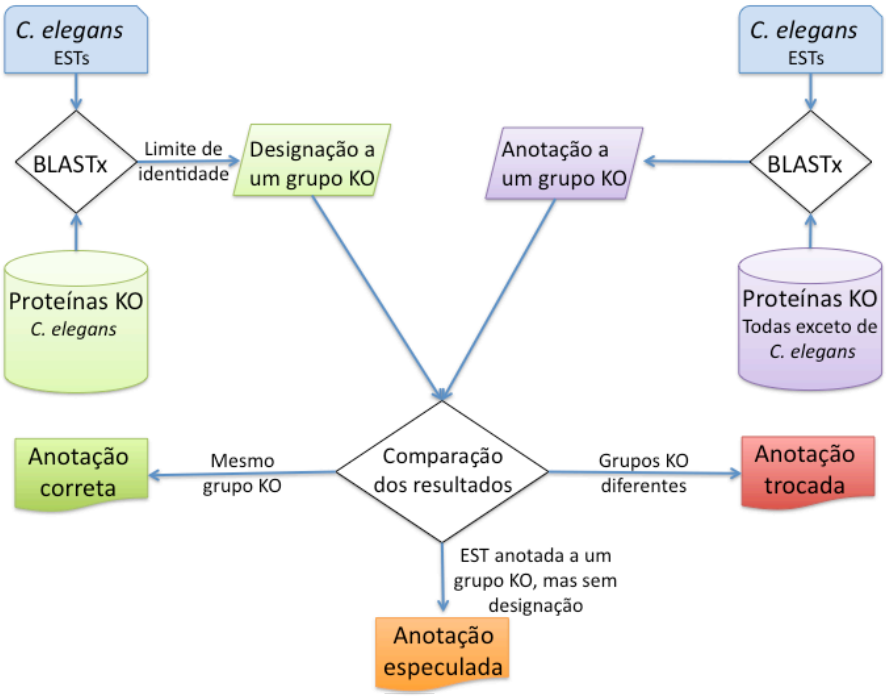

Figura 6 - Teste de anotação para avaliação de uma base de dados. A comparação dos grupos de ortólogos aos quais as ESTs foram associadas em cada uma das etapas leva à classificação da anotação em correta (quando a proteína usada para designação encontra-se no mesmo grupo de ortólogo que a proteína usada para anotação), anotação trocada (quando designação e anotação apontam para grupos diferentes), e anotação especulada (quando temos somente a anotação, sem a designação para comparar).

3.3.1 Designação a proteínas fora do KO

Diante da ausência de entradas no KO capazes de designarem algumas ESTs, buscamos por essas proteínas cognatas executando um BLAST com os mesmos parâmetros para a designação do experimento anterior, mas como base de dados utilizamos todas as entradas de *C. elegans* contidas na base KEGG Genes e UniProtKB.

3.4 INTEGRAÇÃO DE BASES DE DADOS

3.4.1 Edição do COG

O primeiro passo para a integração de COG com UniProtKB é remover as sequências obsoletas, com GI inválido ou sem associação a um identificador UniProt. Para isso seguimos uma linha de procedimentos descrita na Figura 7 e no artigo (Fernandes, Barbosa *et al.*, 2008)

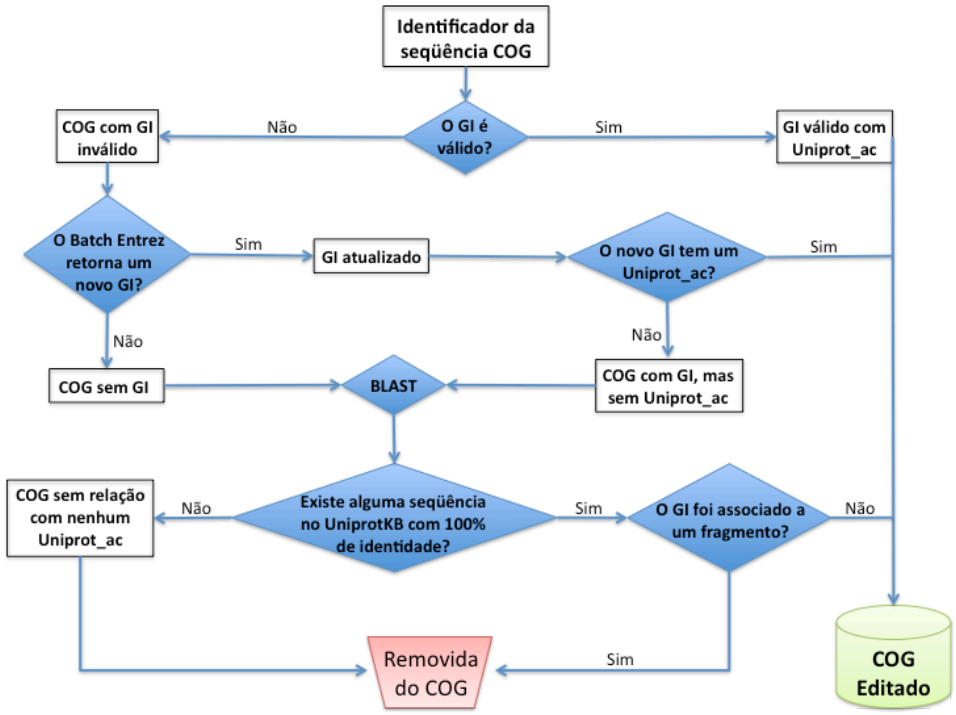

Figura 7 - Procedimentos para criação do COG Editado. Esta etapa tem como objetivo associar cada entrada COG a uma proteína UniProt válida, removendo da base editada as entradas em que essa associação não seja possível.

3.4.2 Criação do UECOG 1.0

A primeira versão da base COG enriquecida foi feita de acordo com a metodologia descrita por nosso grupo no artigo (Fernandes, Barbosa *et al.*, 2008). Consiste basicamente em recrutar a um agrupamento COG todas as proteínas procarióticas pertencentes ao mesmo agrupamento UniRef50 que as proteínas recrutadoras, aplicando-se um filtro de tamanho, onde só são recrutadas sequencias com menos de 10% de diferença de tamanho com a recrutadora. Esse recrutamento acontece em um ambiente MySQL onde associamos os identificadores UniProt dessas proteínas aos grupos UniRef50, e a partir daí utilizamos esse

identificador comum para executar o enriquecimento da base.

A validação por SeedLinkage consiste na execução desse programa e observação se todas os membros no novo conjunto de ortólogos agrupam-se de acordo com os critérios do programa.

O PSI-BLAST foi realizado utilizando as proteínas recrutadoras como entrada, e as proteínas recrutadas foram utilizadas com base de dados de busca para o procedimento. Um limite de valor-e de 10^{-5} foi aplicado e o processo foi repetido até a convergência. Proteínas recrutadas que apresentavam pelo menos um alinhamento com alguma recrutadora foram validadas.

3.4.3 Criação do UECOG 2.0

Seguindo o mesmo objetivo de reunir proteínas que compartilhem o mesmo UniRef50 executamos uma busca por homologia através do programa BLAST. Para cada agrupamento UniRef50 realizamos uma busca em que as proteínas recrutadoras eram as sequências a serem identificadas (*query*) e a base de dados para a busca eram as demais proteínas pertencentes ao agrupamento UniRef50. Os parâmetros utilizados foram uma restrição do valor "e" em 10^{-10} e um alinhamento que cobrisse a sequência recrutadora em pelo menos 50%. Todas as proteínas da base de dados que satisfizessem esses requisitos foram usadas para enriquecer o COG ao qual aquele recrutador pertence.

3.4.4 Criação do UEKO

A integração da base de dados KO com o UniProtKB deu-se seguindo três critérios.

O primeiro deles segue o recrutamento básico de proteínas que compartilhem o mesmo grupo UniRef50, como citado no artigo (Fernandes, Barbosa *et al.*, 2008) e descrito na Figura 8.

O segundo critério obedece um simples alinhamento entre as sequências recrutadoras e recrutadas (alinhamento determinado pelo programa BLAST com limite de valor-e para 10^{-10}). O terceiro critério utiliza o alinhamento gerado pelo BLAST e além disso exige uma cobertura de pelo menos 50% da sequência da proteína recrutadora.

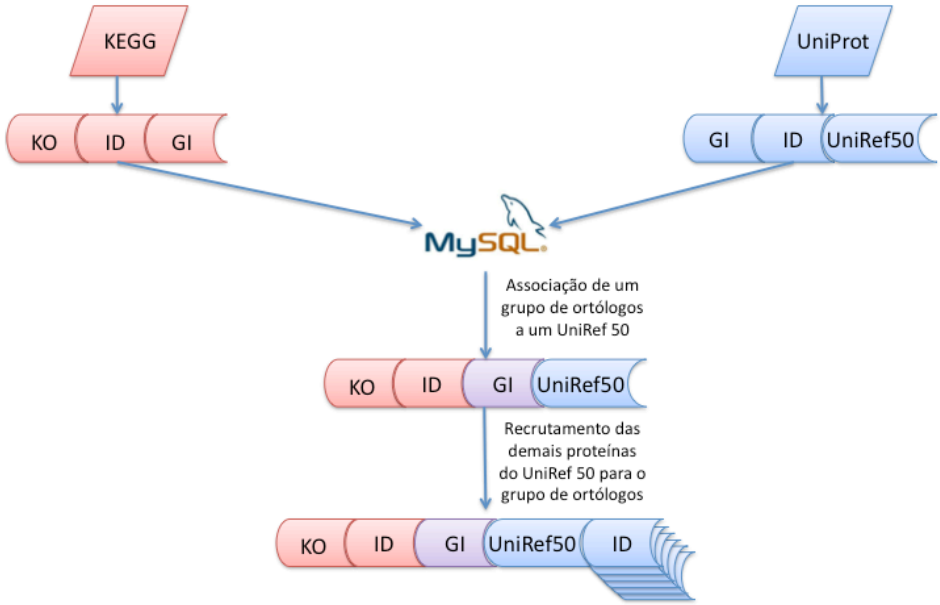

Figura 8 - Procedimento para recrutar proteínas do UniRef50 para o KO. Através de um identificador comum, o GI, é possível a integração entre KEGG e UniProt, que culmina no enriquecimento de sequências.

3.5 ANÁLISE METAGENÔMICA

3.5.1 Anotação filogenética

Os arquivos em formato FASTA contendo as sequências de cada amostra metagenômica foram submetidas a uma busca utilizando o BLAST contra a base de dados NT do NCBI. O resultado da execução do programa foi então carregado em uma ferramenta de análises chamada MEGAN (Huson, Auch *et al.*, 2007). Essa ferramenta identifica a associação entre a sequência e um determinado organismo, ou a associa ao grupo ancestral mais próximo caso o organismo não possa ser precisamente predito. Uma tabela contendo a quantidade de sequências associadas a cada clado é gerada e utilizada para futuras análises. Utilizamos as entradas associadas a gêneros e, quando as sequências eram designadas a espécies, agregávamos essas contagens respectivo gênero.

3.5.2 Anotação funcional

As mesmas sequências citadas anteriormente são montadas de modo a formarem *contigs*. O programa utilizado foi o TGICL (Pertea, Huang *et al.*, 2003) que faz uma otimização com MegaBLAST da montagem do CAP3 (Huang e Madan, 1999).

Após a montagem os *contigs* e *singletons* são submetidos a uma predição de genes. Essa predição é feita pelo programa GeneMark que utiliza de HMM para a busca e utiliza como referência amostras de diversos organismos procariotos, arquivo presente junto com o pacote do programa. Sequências de aminoácidos preditas são então submetidas a uma busca utilizando o BLAST em bases de dados como KO e UEKO (limiar de valor-e $1x10^{-10}$, filtro de baixa complexidade desligado). O resultado com melhor escore do BLAST é então considerado como sendo a informação propagada para a amostra metagenômica.

Para a análise da representatividade de vias, módulos e categorias funcionais a quantidade de sequências assinadas a uma essas categorias (via, por exemplo) foi normalizada dividindo-se pela quantidade de grupos de ortólogos que compõem a categoria.

3.6 ANÁLISES COMPLEMENTARES

3.6.1 Análise de componentes principais (PCA)

Gráficos PCA foram gerados utilizando uma matriz contendo a proporção de cada clado (análise filogenética) ou grupo de ortólogo (análise funcional) em cada amostra metagenômica. Essa matriz é composta com as observações (clado ou KO) nas linhas e as amostras distribuídas nas colunas. Essas matrizes foram carregadas no programa R para os cálculos das matrizes e geração da figura.

3.6.2 Cladogramas

Para a criação de cladogramas foi inicialmente executado o programa ClustalW2 utilizando como entrada um arquivo em formato FASTA contendo as sequências da proteínas a serem comparadas. O cladograma gerado pelo programa foi visualizada na ferramenta iTOL.

Para comparar a troca de KO na análise metagenômicas utilizamos um algoritmo de alinhamento que trata as inserções corretamente e evita as superestimativas do número de eventos de deleção. Para isso o alinhamento múltiplo foi realizado com o programa Prank (Loytynoja e Goldman, 2010) disponível em http://www.ebi.ac.uk/goldman-srv/webPRANK/, o resultado foi submetido aos programas do pacote Phylip 3.67 Protdist e Neighbor, serviço disponível em http://mobyle.pasteur.fr/ obtendo-se um arquivo contendo a informação para a construção do cladograma, o qual foi gerado no serviço iTOL mencionado acima.

3.6.3 Análise do último ancestral comum

Para a análise de ancestral comum das proteínas dos testes de anotação utilizando KO e UEKO selecionamos os identificadores taxonômicos (Sayers, Barrett *et al.*, 2011) das proteínas KO e UEKO utilizadas para a anotação de cada EST. Esses dois identificadores foram utilizados como entrada para um código em PERL que acessa um serviço *WEB* desenvolvido por Henrique Velloso, do Laboratório de Biodados da UFMG, que retorna o ancestral comum daquele conjunto de identificadores usados como entrada. Selecionamos então a profundidade cladística desse ancestral comum (ex.: classe, ordem). O resultado indica o clado ancestral mais próximo que agrupa as proteínas anotadoras do KO e do UEKO.

A mesma consulta ao serviço *WEB* foi feita para a análise das anotações às sequências metagenômicas.

3.6.4 Identificação dos enterotipos

A análise de amostras metagenômicas que resultou na identificação dos enterotipos foi realizada, em sua grande parte, durante estágio no EMBL, em Heidelberg, Alemanha. Créditos são dados a Mani Arumugam e Daniel Mende que auxiliaram na anotação filogenética, e a Takuji Yamada e Jeroen Raes que participaram da análise funcional.

3.6.5 Obtenção das sequências

As amostras européias foram selecionadas, coletadas, preparadas e sequenciadas de

acordo com a descrição no artigo publicado (Arumugam, Raes *et al.*, 2011). As sequências japonesas, como citadas anteriormente, foram publicadas (Kurokawa, Itoh *et al.*, 2007) e disponibilizadas. As amostras dos indivíduos americanos foram obtidas através do artigo (Gill, Pop *et al.*, 2006).

Para complementar as análises, amostras de 154 americanos foram obtidas em http://gordonlab.wustl.edu/NatureTwins_2008/V2.fasta.gz, amostras essas de sequências de 16S rDNA publicadas (Turnbaugh, Hamady *et al.*, 2009).

Dados dos 85 dinamarqueses sequenciados em equipamento Illumina foram obtidos por um projeto anterior também já publicado (Qin, Li *et al.*, 2010)

3.6.6 Processamento e montagem dos metagenomas

A anotação filogenética de cada amostra deu-se usando o programa SMASH (Arumugam, Harrington *et al.*, 2010), com os parâmetros de "E=1e-20 Z=4000000000 B=5" que são passados para o BLAST limitar o valor-e, simular o tamanho da base de dados e selecionar o número de resultados relevantes, respectivamente. Para o mapeamento foram utilizados 1511 genomas como referências, obtidos no NCBI, através do Projeto Microbioma Humano (Nelson, Weinstock *et al.*, 2010), e fornecidos pelo consórcio MetaHIT, sendo adquiridos em http://www.sanger.ac.uk/resources/downloads/bacteria/metahit/.

Os genes de rRNA 16S foram identificadas utilizando um algorítmo baseado em HMM (Huang, Gilna *et al.*, 2009) e assinados a um nível taxonômico utilizando o RDP Classifier (Wang, Garrity *et al.*, 2007).

A anotação funcional desse grupo de amostras foi realizada pelo programa SMASH. A montagem das sequências consenso foi feita internamente pelo programa ARACHNE (Batzoglou, Jaffe *et al.*, 2002) e a predição de genes foi igualmente realizada pelo programa GeneMark, citado anteriormente.

3.6.7 Definição dos grupos e análises estatísticas

Para a definição dos grupos que formam os enterotipos utilizamos um algoritmo de agrupamento chamado PAM, que agrupa os perfis de representatividade. PAM é derivado do algoritmo "k-means", mas tem a vantagem de suportar medidas de distância arbitrárias e é

mais robusto que o "k-means" (Kaufman e J. Rousseeuw, 2005). Para a definição do tamanho dos grupos utilizamos o índice de Calinski-Harabasz, indicado para identificação da quantidade de grupos em uma amostra (Calinski e Harabasz, 1974) (Milligan e Cooper, 1985).

A delimitação dos grupos na análise de componentes principais é feita por um pacote chamado "ade4" disponível para o programa R. Para esta análise entre classes necessitamos somente da análise do componente principal das amostras e um conjunto de classes (grupos definidos pelo PAM); ela nos permite encontrar os componentes principais baseados no centro de gravidade de cada grupo e a evidenciar as diferenças entre eles, além de ligar cada amostra a um grupo.

A rede de correlação entre os gêneros responsáveis pela criação dos enterotipos foi feito pelo pacote "network" e utiliza a correlação de Spearman para fazer as ligações.

A definição de sub e super-representação de gêneros e funções foi dada por um teste exato de Fish em que consideramos um valor-p de no máximo 0,05 para tomarmos tais conclusões.

Todas essas análises foram feitas no programa R e os pacotes usados estão disponíveis em seu repositório.

4 RESULTADOS

4.1 CONSTRUÇÃO DE UMA BASE DE DADOS LOCAL

4.1.1 KEGG

Arquivos obtidos do servidor FTP do KEGG (ver Material e Métodos), referentes à versão 56 datada de 30 de setembro de 2010, foram processados e sua informação foi armazenada em um banco de dados MySQL. Os produtos de 5.636.919 genes oriundos de 1.315 genomas diferentes foram identificados nesses arquivos. Desse total apenas 2.122.701 proteínas estavam associadas a pelo menos um de 13.581 grupos de ortólogos ("KEGG Orthology group, ou KO), cerca de 37,66% delas. Alguns organismos como *Haemophilus influenzae Rd KW20* tem uma boa parte de sua informação organizada em KO, com 1.299 dos seus 1.657 genes (78,39%) contidos na base associados a um grupo de ortólogos. Por outro lado, organismos modelo para estudos como *Branchiostoma floridae* - o anfioxo - e *Oryza sativa* - o arroz – têm apenas 11,94% das 28634 entradas, e 13,15% de 26937 proteínas nos grupos KO, respectivamente.

Além das informações das proteínas e seus grupos de ortólogos foram associados também dados de 307 vias metabólicas e não metabólicas, e 36 categorias funcionais para ampliar as análises feitas com a base de dados a níveis sistêmicos.

4.1.2 UniProtKB

Através do servidor FTP do UniProt foram obtidos os dados relativos à base de dados UniProtKB em sua versão 2010_10, referente ao mês de outubro de 2010. Informações sobre 521.016 proteínas que estavam contidas na base do Swiss-Prot e 12.098.541 na base trEMBL, foram armazenadas. Essas informações gerais consistem na integridade da proteína – se ela é completa ou apenas um fragmento, tipo de evidência sobre a existência de cada entrada, assim como tamanho da sequência de aminoácidos, GI's associados à entrada UniProt_ac, identificados taxonômico do organismo ao qual pertence e seu grupo UniRef50. Para 1.528.387 proteínas (12,11%) foi possível obter também o número EC, o que permite uma melhor identificação da função desempenhada.

Esse conjunto de dados foi armazenado em tabelas de um banco MySQL (ver Material e Métodos), devidamente organizadas e indexadas para a otimização do processo de integração.

4.2 TESTE DE ANOTAÇÃO USANDO KO COM BASE DE DADOS

Um estudo para avaliar a capacidade de anotação automatizada da base de dados KO foi realizado em um trabalho prévio (Fernandes, Mudado *et al.*, 2008). Nesse trabalho avaliamos a base de dados utilizando ESTs de quatro organismos modelo (*Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Homo sapiens*) e observamos a influência da ausência de informação dos organismos nas especulação das anotações. Após três anos realizamos o mesmo experimento para observar se as atualizações periódicas da base de dados KO teriam resolvidos esse problema. Para isso, um total de 171.372 ESTs de *Caenorhabditis elegans* obtidas do NCBI foram usadas como conjunto para consulta usando o BLAST em uma base de dados formada pelas sequências de proteínas contidas no KO. Essas ESTs são oriundas de diversas bibliotecas de vermes de ambos os sexos e todos os estágios de desenvolvimento, a fim de reproduzir o experimento do artigo referido acima com a versão mais recente do KO.

Ao consultarmos o alinhamento das ESTs com uma base de dados contendo somente proteínas de *C. elegans* do KO obtivemos 54.747 associações, as quais nós chamamos de designação, dados os critérios de corte (ver Material e Métodos). Por outro lado, ao utilizarmos toda a base do KO, exceto as proteínas de *C. elegans*, obtivemos 70.036 ESTs anotadas a grupos KO. Portanto, nem todas as ESTs anotadas haviam sido designadas a proteínas do verme.

Para essas ESTs que foram anotadas, mas não foram designadas à proteína cognata no próprio *C. elegans*, atribuímos o rótulo de especuladas, e somaram 19.830. As demais entradas com designação e anotação a um grupo de ortólogo foram comparadas, e observamos 2.562 ESTs onde o KO designado era diferente do KO anotado, caracterizando uma troca. Porém, 47.644 ESTs (94,9% das designadas) mantiveram o mesmo KO ao longo dos dois experimentos. A distribuição dessas classes de anotação pode ser visualizadas na Figura 9 - A. Como visto, a grande quantidade de especulação mostrada no artigo publicado ainda se mantém.

4.2.1 Designação das EST com anotação especulada

Diante do grande número de especulações a busca por uma designação a proteínas de *C. elegans* fora da base de dados do KO apresentou-se como uma possível solução, caracterizando-se assim a ausência da informação de *C. elegans* nos grupos de ortólogos, fato mostrado na publicação anteriormente mencionada. Nessa versão do experimento a designação das ESTs a proteínas fora do KO solucionou 16.746 casos de especulação. As proteínas do KEGG extra-KO com mais designações estão expostas na Tabela 2, as quais são associadas pelo KEGG aos identificadores UniProtKB mostrados na tabela. Marcadas com asterisco estão proteínas do verme que são incorporadas à base KO pelo procedimento de enriquecimento da mesma, mostrado adiante.

A distribuição das classes de anotação após considerar a designação fora do KO está na Figura 9 - B (área verde claro). Aparentemente proteínas de *C. elegans* teriam um potencial para serem incorporadas à base anotadora, embora não seja possível, automaticamente, verificar se a designação a proteínas extra-KO e anotação são equivalentes.

A origem dessas proteínas às quais as ESTs com anotação especulada são designadas fora do KO é diversa. Algumas são entradas exclusivas da base UniProtKB, algumas são exclusividade da base KEGG, mas a grande parte das designações foi feita a entradas compartilhadas pelas duas bases de dados. Essa distribuição pode ser vista na Figura 9 - C.

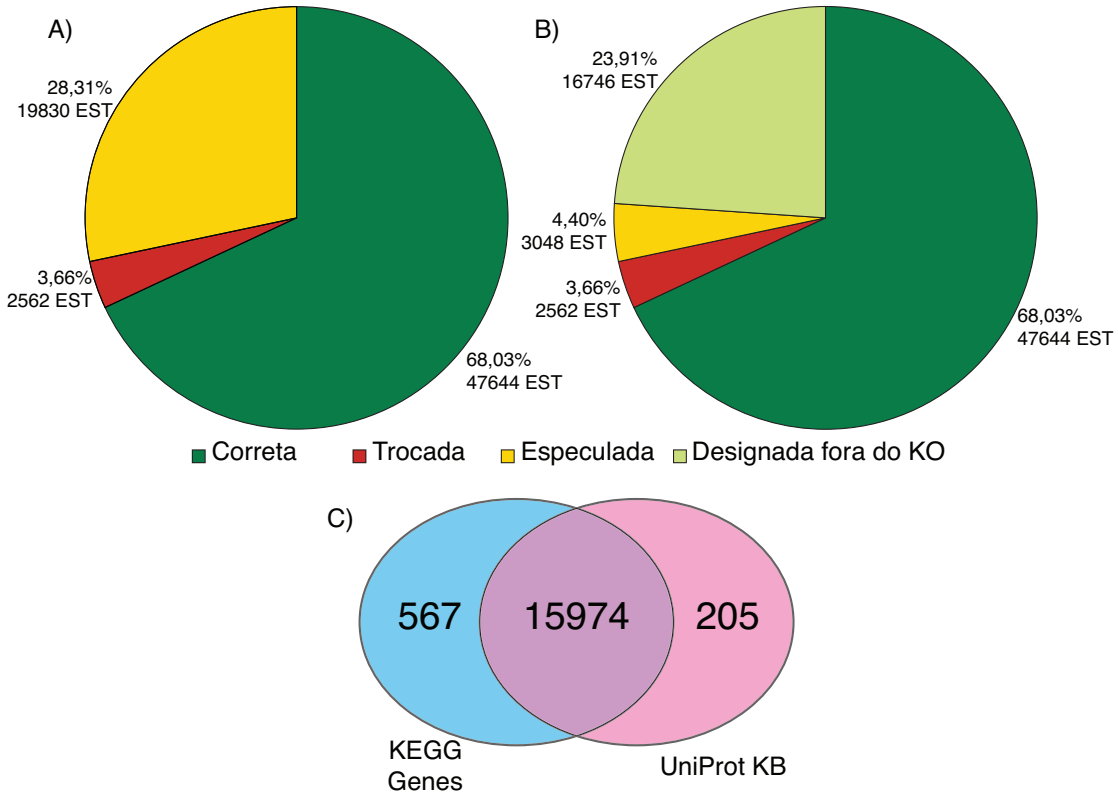

Figura 9 - Teste de anotação de EST de *C. elegans* com KO. A) Um total de 171.372 de ESTs foram submetidas a uma busca para serem designadas a proteínas de *C. elegans* e/ou anotadas com as demais proteínas da base, resultando em anotação correta, trocada ou especulada, como indicado. B) Mesmo experimento, mas alinhando especuladas com proteínas de *C. elegans* complementares oriundas de KEGG Genes e UniProt KB (Designada fora do KO). C) Número de ESTs inicialmente especuladas que são designadas a proteínas de *C. elegans* de fora da base KO.

Tabela 2 - Principais proteínas para as quais foram designadas ESTs com anotação especulada.

Identificador no KEGG	Identificador no UniProtKB	Ocorrências	Descrição segundo o KEGG
cel:F54C1.7	P91328	234	pat-10; Paralysed Arrest at Two-fold family member (pat-10)
cel:Y106G6H.2	Q9U302 Q7K798* Q7K797	221	pab-1; PolyA Binding protein family member (pab-1)
cel:JC8.3	P61866*	213	rpl-12; Ribosomal Protein, Large subunit family member (rpl-12)
cel:Y25C1A.7	Q2Z1N6 Q2Z1N7 Q9TYM1	191	hypothetical protein
cel:C05E4.9	Q10663 Q8IA71	180	gei-7; GEX Interacting protein family member (gei-7)
cel:F13B10.2	P50880* Q6BEU5 Q6BEU6	140	rpl-3; Ribosomal Protein, Large subunit family member (rpl-3)
cel:C55B7.4	Q966M3* Q8IAB6	121	acdh-1; Acyl CoA DeHydrogenase family member (acdh-1)
cel:C25B8.3	P43510* Q8MQC6 A7LPD1	113	cpr-6; Cysteine PRotease related family member (cpr-6)
cel:R11A5.4	O02286 Q7JKI1* Q7JKI3 Q7JKI2	98	hypothetical protein
cel:ZK622.3	Q23552 Q95PW7 Q8IFX3 Q86NB3	95	pmt-1; Phosphoethanolamine MethyTransferase family member (pmt-1)
cel:Y39B6A.20	Q9TVS4	91	asp-1; ASpartyl Protease family member (asp-1)
cel:F10C1.7	Q19286	84	ifb-2; Intermediate Filament, B family member (ifb-2)
cel:Y82E9BR.3	Q9BKS0*	82	hypothetical protein
cel:F58B3.1	Q20964	82	lys-4; LYSozyme family member (lys-4)
cel:C44B7.10	Q18599	82	hypothetical protein

* incorporadas ao UEKO (ver adiante)

4.3 UECOG

Uma de nossas iniciativas visando enriquecer em informações bases de dados secundárias foi inicialmente realizada com a base COG (Fernandes, Barbosa *et al.*, 2008).

Após os procedimentos descritos em Material e Métodos, obtivemos uma base COG com algumas modificações. Após a remoção de GIs inválidos, entradas que correspondiam a fragmentos de proteínas, e sequências de eucariotos, restaram 124.369 proteínas de 63 organismos, em comparação com as 144.320 entradas da base original.

Essa base que chamamos de COG Editado foi usada para recrutar mais 837.356 proteínas de 3.414 organismos diferentes utilizando como referência os grupos UniRef50.

Como o COG Editado é composto apenas por procariotos, fizemos uma análise em diferentes clados proeminentes e para cada etapa do processo de construção. As quantidades de proteínas e organismos estão sumarizadas na Tabela 3. A base UECOG contém, portanto, proteínas provenientes de mais genomas, ao lado de uma quantidade maior de proteínas, auxiliando estudos comparativos por oferecer maior quantidade de proteínas relacionadas já pré-agrupadas.

Tabela 3 - Enriquecimento do UECOG

Clado	COG		COG Editado		COG Enriquecido		*Aum.
	Genomas	Proteínas	Genomas	Proteínas	Genomas	Proteínas	
COG	66	144320	n/a	n/a	n/a	n/a	n/a
Procariotos	63	137122	63	124369	3477	961725	7,01
Archaea	13	22374	13	21310	248	49836	2,23
Bactéria	53	114748	50	103059	3229	911889	7,95
Actinobacteria	4	9391	4	6736	391	65871	7,01
Firmicutes	12	20921	12	19961	747	184403	8,81
Proteobacteria	24	67737	24	60741	1594	592000	8,74
Outras Bactérias	14	16699	10	15621	497	69615	4,17

n/a = não aplicável
* Número de vezes que a base aumentou em nível protéico.

Após os procedimentos que definiram o recrutamento de membros UniRef50 por elementos recrutadores do COG Editado, alguns testes foram realizados para a validação do enriquecimento. Um grupo de ortólogos UECOG0151 (phosphoribosylamine-glycine ligase), foi submetido à validação por Seed Linkage e PSI-BLAST, como descrito em Material e Métodos. O resultado geral e por clados pode ser visualizado na Tabela 4. Como não era possível naquela época realizar a validação global da base, um filtro de diferença de tamanho (calibrado para aceitar no máximo 10% de diferença) entre recrutador e recrutado foi incorporado, e isto foi suficiente para eliminar falso-positivos (29, que não eram validados por nenhum método; não mostrado). A seleção por tamanho incorre em perda de informação; para Procariotos (compare a linha Procariotos com Procariotos*) são perdidos 12 validados por PSI-BLAST e 10 validados por Seed Linkage, todavia a especificidade sobe para 100% e

99%, respectivamente.

A base de dados UECOG está disponível em http://biodados.icb.ufmg.br/uecog. Na página é possível ao usuário adquirir as sequências da base de acordo com a sua demanda, podendo ser obtidas todas as proteínas responsáveis por uma dada função, de um grupo de ortólogos ou de um grupo de organismos. Uma visualização da interface da página pode ser vista na Figura 10.

Tabela 4 - Validação do UECOG0151 por PSI-BLAST e Seed Linkage.

Clado	Recrutador	Recrutado	PSI-BLAST	Seed Linkage
Procariotos	51	501	470 (94%)	463 (92%)
Procariotos*	51	459	458 (99,9%)	453 (99%)
Archaea	12	30	30 (100%)	29 (97%)
Bactéria	41	471	438 (93%)	434 (92%)
Actinobactéria	4	41	38 (93%)	40 (98%)
Firmicutes	8	93	92 (99%)	92 (99%)
Proteobactérias	21	313	279 (89%)	282 (90%)
Outras Bactérias	6	24	20 (83%)	20 (83%)

* Usando 10% de limite de tamanho no recrutamento

UE-COG - UniRef-Enriched COG Database

UE-COG division:

Legend (click on numbers to download fastas):

- Pk = prokaryote (all)
- A = Archea
- B = Bacteria
- Ac = Actinobacteria
- Fm = Firmicutes
- Pb = Proteobacteria
- Ot = other bacteria

Category	Function	Number of sequences per division (number of enriched sequences inside parenthesis)						
		Pk	*A*	*B*	*Ac*	*Fm*	*Pb*	*Ot*
D	Cell cycle control, cell division, chromosome partitioning	1439 (11183)	205 (453)	1234 (10730)	123 (847)	234 (2433)	653 (6646)	224 (804)
M	Cell wall/membrane/envelope biogenesis	7688 (56732)	676 (1507)	7012 (55225)	414 (2964)	1113 (10024)	4380 (37760)	1105 (4477)
N	Cell motility	2660 (18867)	245 (488)	2415 (18379)	101 (536)	319 (1812)	1686 (15121)	309 (910)
O	Posttranslational modification, protein turnover, chaperones	5500 (47223)	880 (2265)	4620 (44958)	355 (2815)	678 (7365)	2818 (30625)	769 (4153)

Figura 10 - Interface de utilização do serviço UECOG. Em cada um dos links o usuário pode obter a versão FASTA das proteínas.

4.3.1 UECOG 2.0

Um recente enriquecimento da base COG foi feito considerando a homologia inferida pelo BLAST entre a proteína recrutadora e recrutadas e, subsequentemente, a cobertura da proteína recrutadora pelo alinhamento, conforme descrito em Material e Métodos. Este procedimento é utilizado atualmente para a produção da base UEKO (ver adiante). O resultado final mostra que esse procedimento é eficaz para o recrutamento, eliminando a necessidade de um filtro de diferença de tamanho entre recrutador e recrutados. Na publicação original do UECOG (Fernandes, Barbosa *et al.*, 2008) estimávamos que recrutadas pudessem divergir em até 30% do tamanho da recrutadora, todavia o valor de 10% era usado como medida de segurança.

Ao final do processo contamos com uma base de dados com 2.450.485 entradas, oriundas de 5.748 organismos distintos.

O enriquecimento das bases de dados UECOG 1.0 e UECOG 2.0 foi medido para cada grupo de ortólogos. Evidentemente, dada a época em que foram realizados, o enriquecimento mais recente se beneficia do crescimento da base UniProtKB. A maioria dos grupos foi aumentada com alguma nova sequência, como pode ser visualizado na Figura 11 - A. Uma análise comparativa entre as duas versões pode ser vista na Figura 11 - B. Nela observamos que alguns grupos diminuíram de tamanho apesar do aumento geral da base de dados, isso se deve tanto ao procedimento de recrutamento que foi diferente, quanto à realocação de entradas em grupos UniRef50 diferentes pelo UniProt.

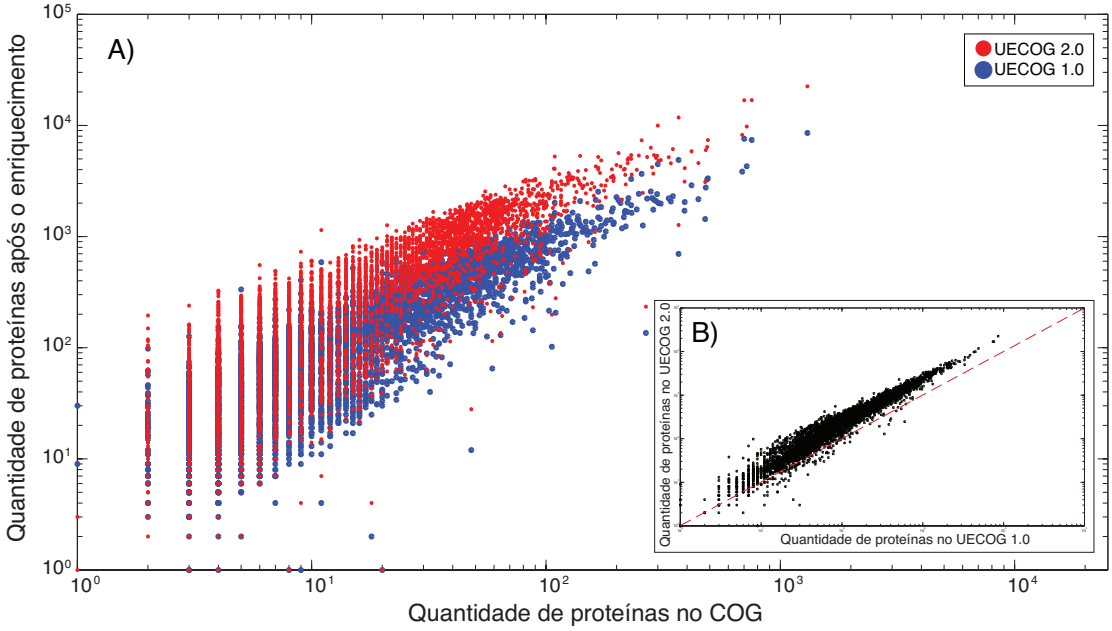

Figura 11 – Enriquecimento de agrupamentos KO em versões do UECOG. A) Número de sequencias em cada agrupamento original (abscissa) versus enriquecido nas versões 1.0 e 2.0 de UECOG (ordenada). B) Número de sequencias em cada agrupamento UECOG 1.0 (abscissa) versus UECOG 2.0 (ordenada).

4.4 INTEGRAÇÃO UNIPROTKB E KEGG

Uma vez que temos as informações das bases de dados em um banco de dados local, a integração dá-se inicialmente pela comparação entre as entradas do UniProt e do KEGG. Essa comparação é possível pela associação de cada entrada ao seu GI.

Das 5.636.919 proteínas do KEGG Genes, divisão que agrupa todas as sequencias, 5.038.269 tinham um correspondente na base de dados UniProtKB, mostrando que cerca de 89,38% do KEGG é coberto pelas mesmas informações no consórcio UniProt.

A unificação das duas bases de dados em questão aumenta consideravelmente o volume de informação disponível, uma vez que os dados muitas vezes são complementares. Ambas as bases tem informações exclusivas, que não são compartilhadas, como pode ser visualizado na Figura 12. Apesar dessas informações não compartilhadas, a integração entre as duas bases mostra-se viável, uma vez que 1.940.617 (91,67%) das 2.116.996 entradas comuns às duas bases estão classificadas nos grupos de ortólogos do KEGG (KO).

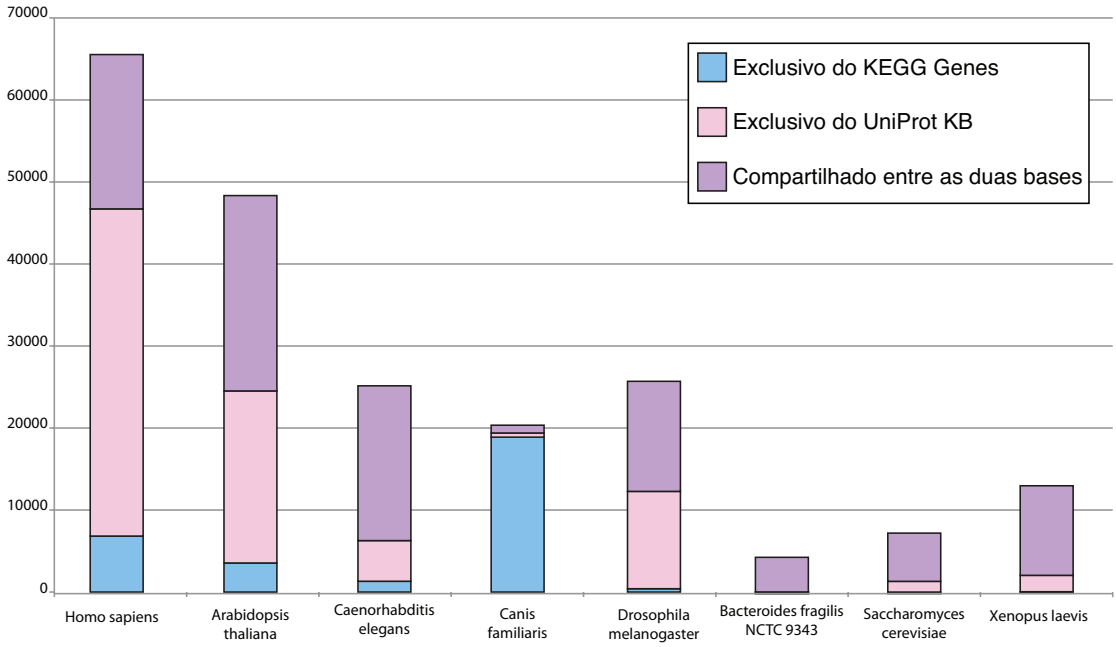

Figura 12 - Disponibilidade de sequencias para diversos organismos em KEGG Genes, UniProt KB ou presentes em ambas as bases. Identificadores GI foram usados para comparação.

4.5 UEKO

Como já havíamos feito para a base COG, procedemos o enriquecimento dos grupos de ortólogos do KEGG, chamados de KO, gerando a base enriquecida UEKO ("UniRef Enriched Kegg Orthology").

Como relatado anteriormente, 1.940.617 proteínas estavam associadas a um KO e a uma entrada UniProt, e consequentemente a um grupo UniRef50. Totalizando temos 436.984 grupos UniRef50 ligados a 13.581 grupos KO.

O primeiro experimento de enriquecimento recrutou todas as proteínas que estivessem no mesmo grupo UniRef50 que uma entrada originalmente no KO, conforme descrito em Material e Métodos. Ao final obtivemos um total de 4.447.538 proteínas de 32.213 organismos diferentes.

Para o segundo experimento, em que o recrutamento é feito com busca de possível homologia, o total de proteínas no UEKO foi de 4.385.157, mas o número de organismos que contribuíram para o enriquecimento foi de apenas 25.108.

O mesmo acontece no terceiro experimento, em que o alinhamento já considerado significativo deve cobrir pelo menos 50% da proteína recrutadora. Isso evita que proteínas contendo domínios promíscuos menores que 50% da recrutadora exerçam o recrutamento, e que proteínas multifuncionais recrutem proteínas para o KO da função imprópria, embora o inverso não seja evitado, de acordo com a postura do KO, que agrupa proteínas multifuncionais nos vários grupos KO onde uni funcionais estão. O número de entradas protéicas diminui para 4.340.702, vindas de 25.035 organismos. Trata-se de um enriquecimento considerável frente às originais 1.940.617 proteínas de 1.315 organismos.

Esse aspecto geral do enriquecimento pode ser visto na Figura 13 em que avaliamos o aumento de cada grupo de ortólogo, em cada um dos experimentos.

Observamos que, ao aplicarmos o filtro de alinhamento através do BLAST, alguns KO tiveram o número final de proteínas reduzido sensivelmente. Os 20 grupos KO que foram mais afetados por esse filtro estão listados na Tabela 5.

Como observado na descrição dos grupos, a grande maioria deles é composto por subunidades de complexos protéicos. O UniRef50 não faz a distinção entre as subunidades, por isso o BLAST torna-se uma ferramenta fundamental para a separação segundo os moldes do KEGG.

Outra redução acontece ao aplicarmos o filtro de cobertura da proteína recrutadora. A redução por este passo foi maior nos grupos descritos na Tabela 6.

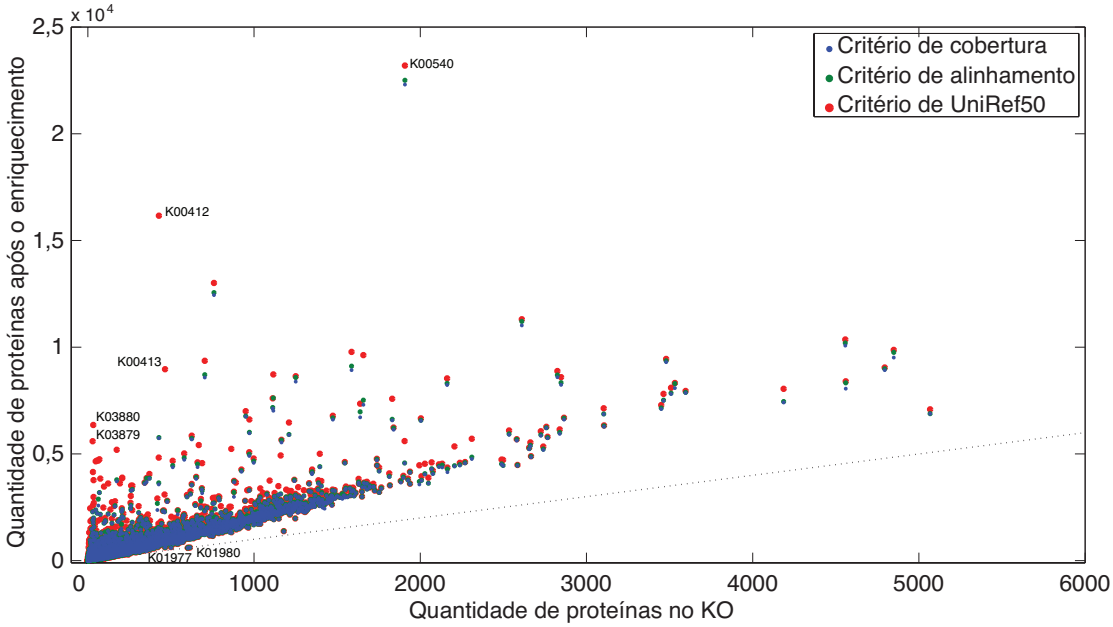

Figura 13 - Análise do tamanho dos agrupamentos UEKO com aplicação de filtros. Está mostrado o tamanho original do agrupamento (abscissa) e enriquecido (ordenada) pelo simples recrutamento com UniRef50, após alinhamento entre recrutadora e recrutadas, e por seleção de alinhamento maior que ou igual a 50% da recrutadora. Alguns agrupamentos KO estão destacados, os quais não são afetados por filtros (K00540) ou são diminuídos sensivelmente (K03879, K03880, K00412 e K00413).

Tabela 5 - **Grupos em que o filtro de alinhamento foi mais efetivo.**

KO	Recrutados pelo UniRef50	Recrutados pelo alinhamento	Filtrados	Descrição
K00412	16160	5771	10389	ubiquinol-cytochrome c reductase cytochrome b subunit [EC:1.10.2.2]
K00413	8974	1433	7541	ubiquinol-cytochrome c reductase cytochrome c1 subunit [EC:1.10.2.2]
K03880	6352	1768	4584	NADH dehydrogenase I subunit 3 [EC:1.6.5.3]
K02276	5413	1915	3498	cytochrome c oxidase subunit III [EC:1.9.3.1]
K02274	5234	1824	3410	cytochrome c oxidase subunit I [EC:1.9.3.1]
K03879	5596	2286	3310	NADH dehydrogenase I subunit 2 [EC:1.6.5.3]
K02262	4691	2218	2473	cytochrome c oxidase subunit III [EC:1.9.3.1]
K06021	3777	1326	2451	phosphate-transporting ATPase [EC:3.6.3.27]
K02126	4659	2377	2282	F-type H+-transporting ATPase subunit a [EC:3.6.3.14]
K03878	4154	1904	2250	NADH dehydrogenase I subunit 1 [EC:1.6.5.3]
K02029	9631	7518	2113	polar amino acid transport system permease protein
K02992	4928	3120	1808	small subunit ribosomal protein S7
K02256	4689	2882	1807	cytochrome c oxidase subunit I [EC:1.9.3.1]
K10004	3856	2068	1788	glutamate/aspartate transport system ATP-binding protein
K01790	3663	1884	1779	dTDP-4-dehydrorhamnose 3,5-epimerase [EC:5.1.3.13]
K10010	3475	1757	1718	cystine transport system ATP-binding protein [EC:3.6.3.-]
K10017	3621	1937	1684	histidine transport system ATP-binding protein [EC:3.6.3.21]
K10008	3839	2169	1670	glutamate transport system ATP-binding protein [EC:3.6.3.-]
K02133	2736	1125	1611	F-type H+-transporting ATPase subunit beta [EC:3.6.3.14]
K01704	3749	2168	1581	3-isopropylmalate/(R)-2-methylmalate dehydratase small subunit

Tabela 6 - Grupos em que o filtro de cobertura foi mais efetivo.

KO	Recrutados pelo alinhamento	Recrutados pela cobertura	Filtrados	Descrição
K01879	2050	1417	633	glycyl-tRNA synthetase beta chain [EC:6.1.1.14]
K07497	7619	7027	592	putative transposase
K04075	2388	1812	576	tRNA(Ile)-lysidine synthase [EC:6.3.4.-]
K12506	907	368	539	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase /
K05685	1104	578	526	macrolide transport system permease protein
K13501	585	68	517	anthranilate synthase / indole-3-glycerol phosphate synthase /
K11808	619	103	516	phosphoribosylaminoimidazole carboxylase [EC:4.1.1.21]
K00955	1089	642	447	
K14048	568	179	389	
K13497	827	455	372	anthranilate synthase/phosphoribosyltransferase [EC:4.1.3.27
K11787	693	347	346	phosphoribosylamine--glycine ligase / phosphoribosylglycinamide
K13797	382	53	329	
K01175	2680	2377	303	
K02256	2882	2583	299	cytochrome c oxidase subunit I [EC:1.9.3.1]
K11755	1219	922	297	phosphoribosyl-ATP pyrophosphohydrolase / phosphoribosyl-AMP
K07496	2796	2509	287	putative transposase
K08282	1472	1190	282	non-specific serine/threonine protein kinase [EC:2.7.11.1]
K00912	1321	1041	280	tetraacyldisaccharide 4'-kinase [EC:2.7.1.130]
K02014	8329	8057	272	iron complex outermembrane recepter protein
K00986	1264	1001	263	RNA-directed DNA polymerase [EC:2.7.7.49]

Grupos como K12506, K13501, K11787, K11755 são compostos por proteínas multifuncionais. Uma vez que o KEGG agrupe corretamente somente proteínas com todas as funções descritas para aquele grupo de ortólogos, o filtro de cobertura seleciona entradas que desempenham todas as funções do seu recrutador, eliminando assim aquelas que contenham pelo menos um domínio conservado. Como dito anteriormente, não almejamos evitar que uma proteína uni funcional opere o recrutamento de uma multifuncional que contenha a função, pois o KO agrupa multifuncionais em todos grupos KO que descrevam as várias funções, por isso utilizamos a cobertura da recrutadora e não a da recrutada.

Como exemplo do efeito dos filtros para o recrutamento analisamos o grupo K00134, GAPDH. Após o enriquecimento com sequências do UniRef50 observamos uma grande quantidade de enolases, transcetolases e triosefosfato isomerases, além das GAPDH. Essas proteínas eram também recrutadas para seus grupos cognatos K01689, K00615 e K01803, respectivamente. Após a aplicação do filtro de homologia entre proteína recrutadora e recrutada obtivemos a eliminação de todas essas entradas erroneamente usadas para o enriquecimento do agrupamento K00134. Com isso um total de 709 transcetolases não foram recrutadas para o K00134 e sim para o K00615, seu respectivo agrupamento. O mesmo acontece com 171 enolases e 72 triosefosfato isomerases.

Outro importante exemplo da sensibilidade e eficiência dos filtros de homologia e de cobertura está associado à entrada A6ZTT5 do UniProt. Essa proteína é uma álcool desidrogenase tipo 4 e ao utilizarmos somente a composição do UniRef50 para o recrutamento ela iria para os grupos K00001 (alcohol dehydrogenase), K00086 (1,3-propanediol dehydrogenase) e K13954. Os filtros aplicados associam essa entrada somente ao grupo K13954, agrupamento específico para as álcool desidrogenase tipo 4.

As entradas que constituem o UEKO podem vir exclusivamente do KEGG Genes (3,3%), UniProtKB (47%) ou podem ser compartilhadas por ambas as bases (49,7%). Alguns organismos são mais influenciados por uma base em específico, devido à disponibilidade de informações sobre tal organismo naquela base. A Figura 14 mostra a composição da base UEKO para alguns organismos, onde evidenciamos a predominância de informações do UniProtKB para *Homo sapiens* e do KEGG Genes para *Canis familiaris*.

A influência da composição das bases de dados originais usadas na elaboração do UEKO exerce grande influência em seu tamanho final. Para ilustrar como ocorrem as atualizações foi realizada uma simulação de enriquecimento utilizando versões antigas do KO e as entradas UniProtKB existentes nas datas de publicação das diferentes versões do KO. Cada versão é atualizada aproximadamente de três em três meses, e a primeira versão do KEGG em que temos agrupamentos KO é a versão 25 datada de 27 de janeiro de 2003. A comparação entre tamanho do KO e do UEKO pra cada versão do KO pode ser visualizado na Figura 15. Note que o enriquecimento só foi feito para a versão atual (55), assim as contagens se basearam na existência dos identificadores KO nas versões anteriores e UniProtKB nessas datas, o que é uma aproximação.

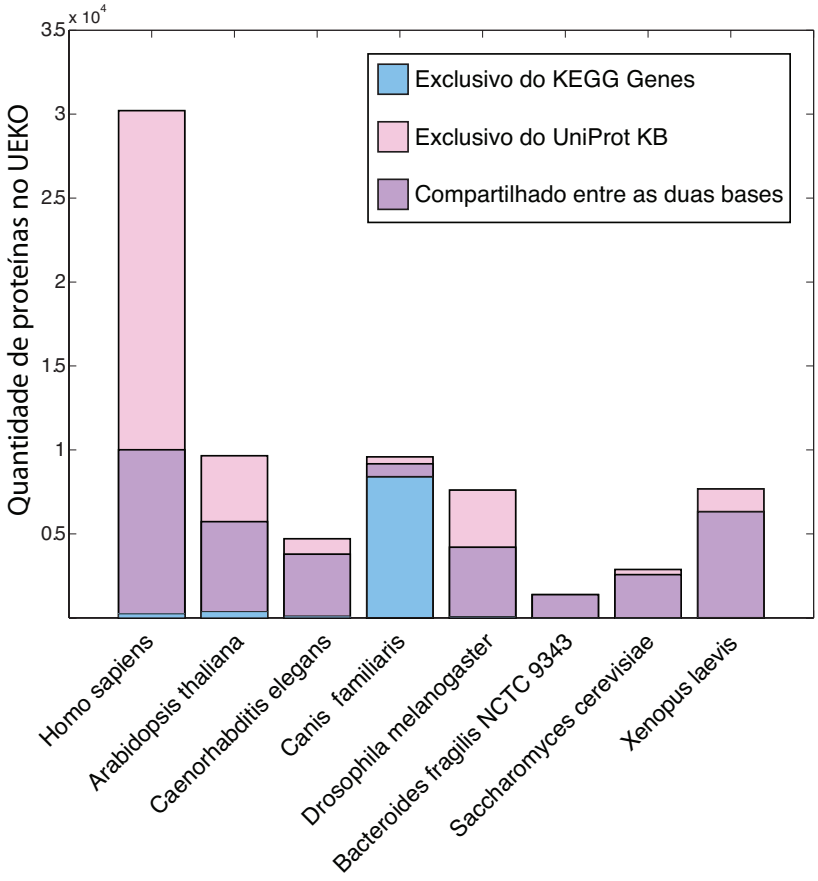

Figura 14 - Composição do UEKO quanto à origem da informação. Para UEKO construído com filtro de cobertura estão mostrados o número de proteínas originais do KEGG Genes, ou adicionadas a partir do UniProtKB ou compartilhadas pelas duas bases de dados.

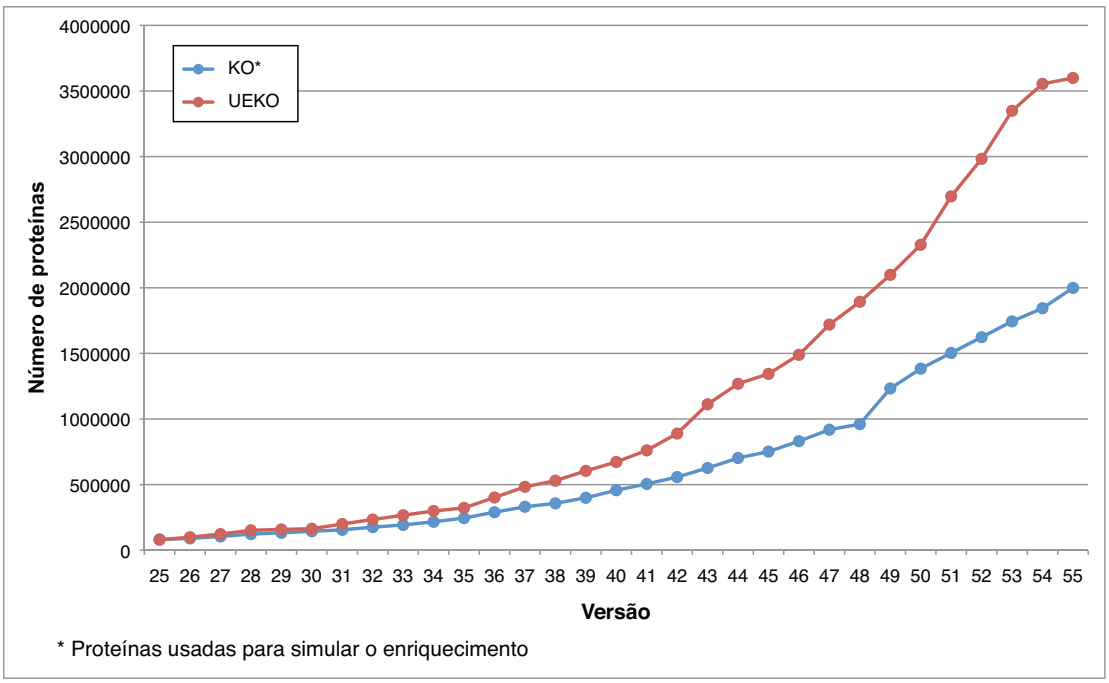

Figura 15 - Simulação do enriquecimento de versões antigas do KO. O tamanho da base UEKO aumenta à medida em que a base referência (KO) aumenta.

4.5.1 Teste de anotação com UEKO

Seguindo os moldes do teste de anotação anterior usamos o UEKO como base de dados para os alinhamentos. A nova referência, com mais que o dobro de entradas, mostra-se eficiente ao diminuir não só a freqüência, mas o número de anotações trocadas e aumentar a freqüência e o número de anotações corretas, porém não soluciona imediatamente as anotações especuladas. A distribuição das classes de anotação pode ser vista na Figura 16.

A busca por designação das ESTs fora da base UEKO revelou que mesmo após o enriquecimento algumas proteínas existentes em *C. elegans* não foram associadas a grupos de ortólogos. Porém, algumas proteínas utilizadas para solucionar a especulação no primeiro teste de anotação, como descritas na Tabela 2, foram efetivamente incorporadas a agrupamentos KO, e estão apresentadas na Tabela 7. Inclusive, a versão atual (56) disponibilizada online já incorpora três dessas proteínas recrutadas pelo UEKO, cel:Y106G6H.2, cel:F13B10.2 e cel:R11A5.4, nos grupos KO indicados por asterisco na Tabela 6.

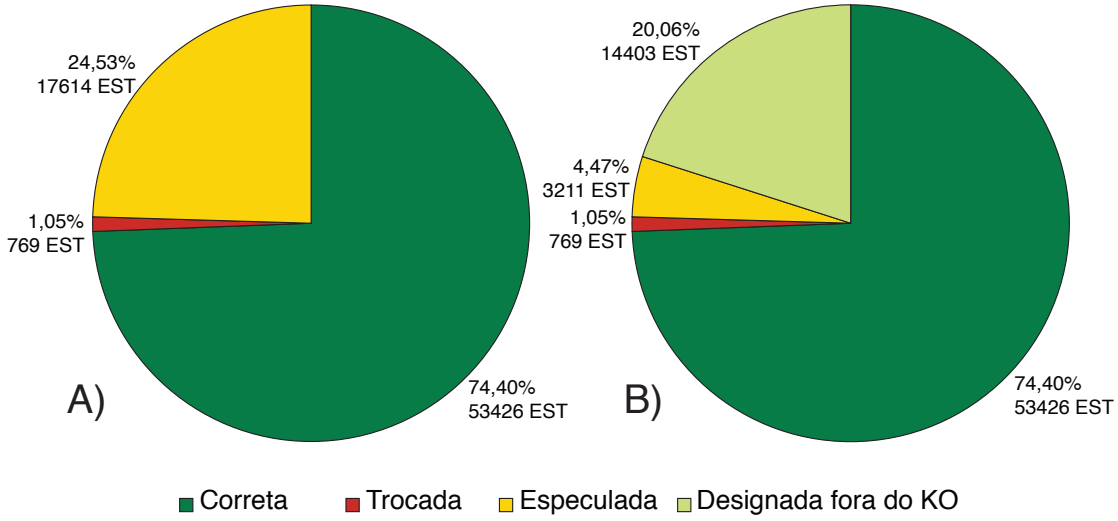

Figura 16 - Teste de anotação de EST de *C. elegans* com UEKO. A) Um total de 171372 ESTs foram assinadas a proteínas de *C. elegans* e/ou anotadas com as demais proteínas da base enriquecida, resultando em anotação correta, trocada ou especulada, como indicado. B) Mesmo experimento, mas alinhando especuladas com proteínas de *C. elegans* complementares oriundas de KEGG Genes e UniProtKB (Assinada fora do UEKO).

Tabela 7 - Proteínas usadas para solucionar a especulação no primeiro teste de anotação e que foram agrupadas em um KO.

Identificador no KEGG	Identificador no UEKO	KO Atribuído	Descrição do KO
cel:Y106G6H.2	Q7K798	K13126*	PABPC; polyadenylate-binding protein
cel:JC8.3	P61866	K02870	RP-L12e, RPL12; large subunit ribosomal protein L12e
cel:F13B10.2	P50880	K02925*	RP-L3e, RPL3; large subunit ribosomal protein L3e
cel:C55B7.4	Q966M3	K00248	E1.3.99.2, bcd; butyryl-CoA dehydrogenase [EC:1.3.99.2]
		K00257	E1.3.99.-
		K09478	ACADSB; short/branched chain acyl-CoA dehydrogenase [EC:1.3.99.12]
cel:C25B8.3	P43510	K01363	CTSB; cathepsin B [EC:3.4.22.1]
cel:R11A5.4	Q7JKI1	K01596*	E4.1.1.32, pckA, PEPCK; phosphoenolpyruvate carboxykinase (GTP) [EC:4.1.1.32]
cel:Y82E9BR.3	Q9BKS0	K02110	ATPF0C, atpE; F-type H+-transporting ATPase subunit c [EC:3.6.3.14]
		K02128	ATPeF0C, ATP5G; F-type H+-transporting ATPase subunit c [EC:3.6.3.14]

* Grupos KO que na versão atual já incorporam a proteína do verme

A comparação das anotações corretas ao usarmos KO ou UEKO observamos que a proteína que anota a EST é diferente ao usarmos as diferentes bases de dados em 12116 casos. Dentre eles, temos 2848 ocorrências de EST que foram anotadas por uma proteína de *Caenorhabditis briggsae*. Entretanto vemos que o UEKO vai além das barreiras do gênero, mostrando 3909 casos em que a proteína anotadora de KO e a de UEKO tem apenas o mesmo reino em comum, coincidindo apenas em clados como Billateria e Celomata. A distribuição das profundidades filogenéticas dos clados compartilhados por anotadores KO e UEKO podem ser vistos na Figura 17. Assim, em casos onde o enriquecimento não produz alteração da anotação, o UEKO tem uma contribuição qualitativa importante.

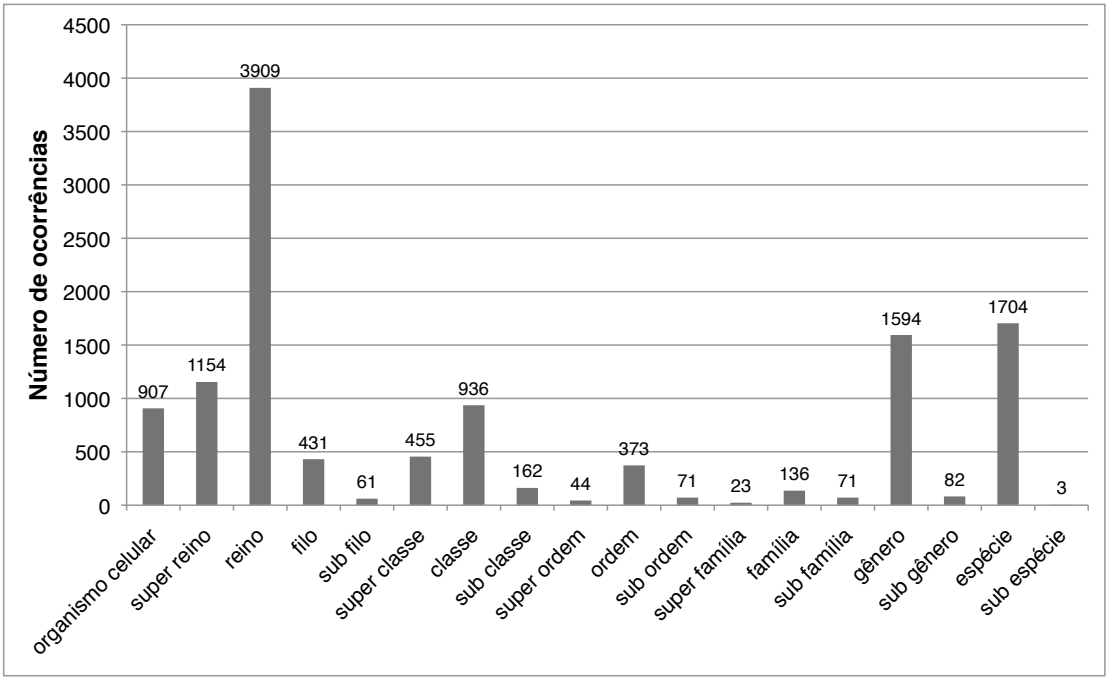

Figura 17 - Profundidade filogenética comparatilhada por proteínas anotadoras de KO e UEKO. O enriquecimento da base KO permite agregar informações de clados tanto próximos quanto distantes dos já presentes na base de dados original. Apenas anotações corretas foram consideradas.

4.5.2 Validação através do número EC

Um total de 1.528.387 entradas do UniProtKB tem pelo menos um número EC associado. Dessas, 1.096.493 (71,74%) estão na base UEKO, sendo 535.388 incorporadas devido ao processo de enriquecimento. Essa informação permitiu a comparação entre os números EC das proteínas recrutadoras e recrutadas. Com isso, obtivemos 1.215.149 (97,9%) de associações que foram validadas como corretas, 15.888 (1,3%) foram classificadas como parcialmente corretas (discordando somente no último algarismo, o que significa mesma função geral), e apenas 10.151 (0,8%) não foram validadas.

Alguns números EC estão associados a entradas do UniProtKB, porém não aparecem no KO. Um exemplo é o EC 3.1.4.41, uma esfingomielina fosfodiesterase D, que tem 214 entradas no UniProtKB, 3 no KEGG Genes, mas nenhuma associada a algum KO. O mesmo acontece com o EC 2.1.1.57, com 109 proteínas no UniProtKB e EC 2.7.7.68, com 89. Assim, uma fonte de construção de novos agrupamentos funcionais pode o trabalho com essas funções não contempladas. Isso mostra que ainda há muita informação a ser organizada nos grupos de ortólogos.

4.6 APLICAÇÃO EM ANÁLISE METAGENÔMICA

Uma análise metagenômica de dados públicos de 13 amostras de microbiota de trato intestinal humano foi utilizada para estimar a contribuição qualitativa e qualitativa do enriquecimento do KO. Os dados sobre idade, gênero, quantidade de sequências, *contigs* e *singlets* encontram-se na Tabela 8.

Tabela 8 - Metadados e informações sobre montagem dos metagenomas.

Nome	Sexo	Idade	Sequências	*Contigs*	*Singletons*
F1-S	M	30 anos	78123	28056	16802
F1-T	F	28 anos	80477	29540	23470
F1-U	*F*	*7 meses*	*80796*	*13396*	*10807*
F2-V	M	37 anos	79846	33684	21427
F2-W	F	36 anos	78670	27004	18560
F2-X	M	3 anos	79773	24018	20816
F2-Y	F	1,5 anos	79357	30234	22498
In-A	M	45 anos	75532	14946	14162
In-B	*M*	*6 meses*	*79972*	*3332*	*4964*
In-D	M	35 anos	80627	23580	27489
In-E	*M*	*3 meses*	*79787*	*11654*	*9192*
In-M	*F*	*4 meses*	*87324*	*18506*	*8936*
In-R	F	24 anos	81346	32704	19644

Infantes marcados itálico.

Cada uma dessas sequências foi anotada a um clado, com o uso do software Megan (Huson, Auch *et al.*, 2007), sendo a fração capaz de ser mapeada apresentada na Tabela 9. A representação dos mais relevantes gêneros mapeados nas amostras é visualizado na Figura 18. Gêneros como *Bifidobacterium*, *Faecalibacterium*, *Bacteroides*, *Eubacterium*, *Parabacteroides* e *Ruminococus* têm uma grande contribuição na maioria das amostras. A contribuição de *Bifidobacterium* é a que mais varia, indo de 2,2% em In-A e 2,8% em In-R (45 e 24 anos, respectivamente), até 92% em In-E e 98% em In-B (sexo masculino, 3 e 6 meses de idade, respectivamente).

Tabela 9 - Fração de sequências mapeadas a um clado taxonômico.

Amostra	Sequências mapeadas
F1-S	51,03%
F1-T	31,04%
F1-U	67%
F2-V	31,82%
F2-W	41,92%
F2-X	26,82%
F2-Y	37,91%
In-A	63,86%
In-B	69,25%
In-D	26,81%
In-E	74,95%
In-M	50,08%
In-R	53,95%

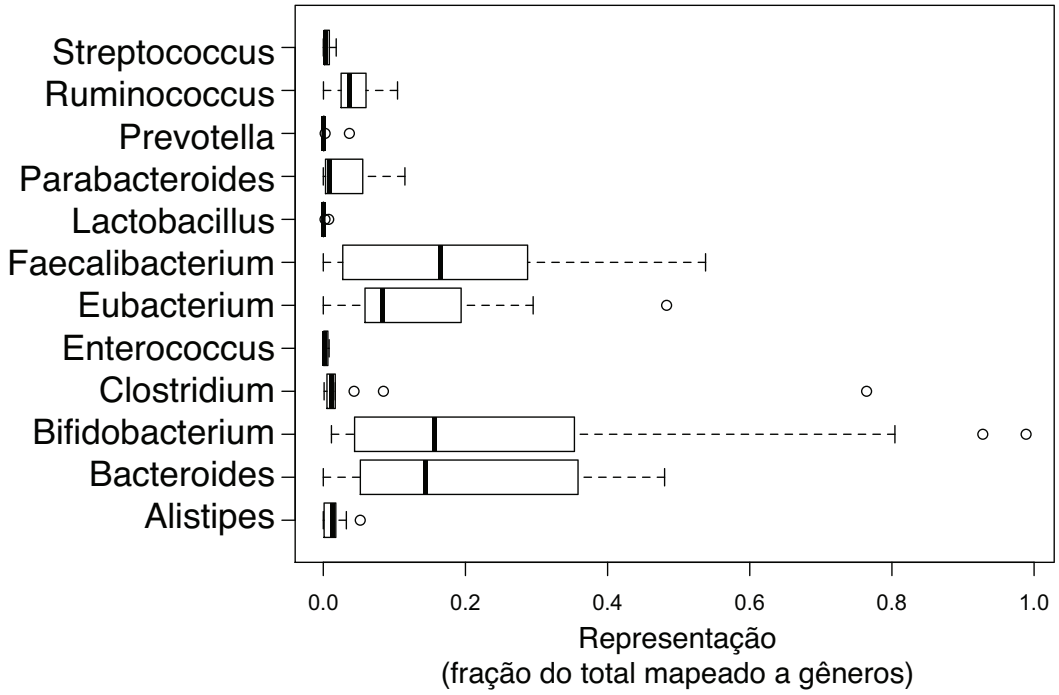

Figura 18 - Contribuição de alguns gêneros para a constituição da microbiota. Representação da fração dos gêneros mais relevantes na microbiota. Estão mostrados os gêneros com representação maior que 0,01.

A análise de componentes principais foi então utilizada para visualizar no espaço a disposição dos indivíduos baseada na composição da sua microbiota, como pode ser visto na Figura 19. Nela podemos observar um isolamento das amostras In-E e In-B guiados principalmente pela prevalência de *Bifidobacterium*. Com isso notam-se dois eixos principais: um eixo horizontal orientado pela contribuição de *Bifidobacterium* nas amostras, guiando principalmente os quatro indivíduos infantes F1-U, In-M, In-E e In-B; e outro eixo vertical guiados por *Bacteroides*, *Eubacterium* e *Faecalibacterium* orientando as demais amostras.

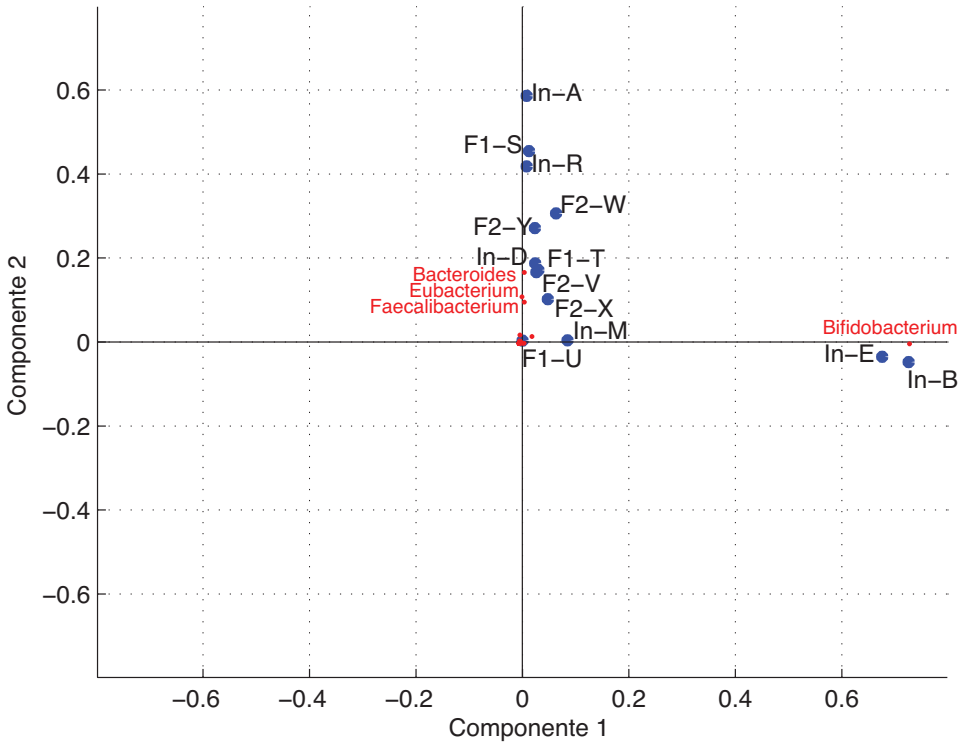

Figura 19 - Análise dos componentes principais mostrando os gêneros orientando a disposição dos indivíduos no espaço. Um eixo de *Bifidobacterium* guia os infantes, enquanto *Bacteróides, Eubacterium* e *Faecalibacterium* orienta os demais indivíduos.

Os *contigs* e os *singletons* são então usados para a predição de regiões codificadoras e consequentemente as proteínas codificadas por tais genes. A quantidade de proteínas preditas assim como seu tamanho médio estão na Tabela 10, e histogramas mostrando o tamanho da proteínas e sua frequência estão na Figura 20.

Tabela 10 - Quantidade e tamanho das proteínas das amostras metagenômicas.

Nome	Proteínas	Média do tamanho (aa)
F1-S	38574	187,38
F1-T	37902	189,39
F1-U	25634	188,42
F2-V	43714	192,32
F2-W	38751	187,90
F2-X	34008	190,11
F2-Y	40924	195,58
In-A	22631	204,51
In-B	9305	228,27
In-D	33856	190,18
In-E	19258	190,94
In-M	29097	201,18
In-R	43688	178,85

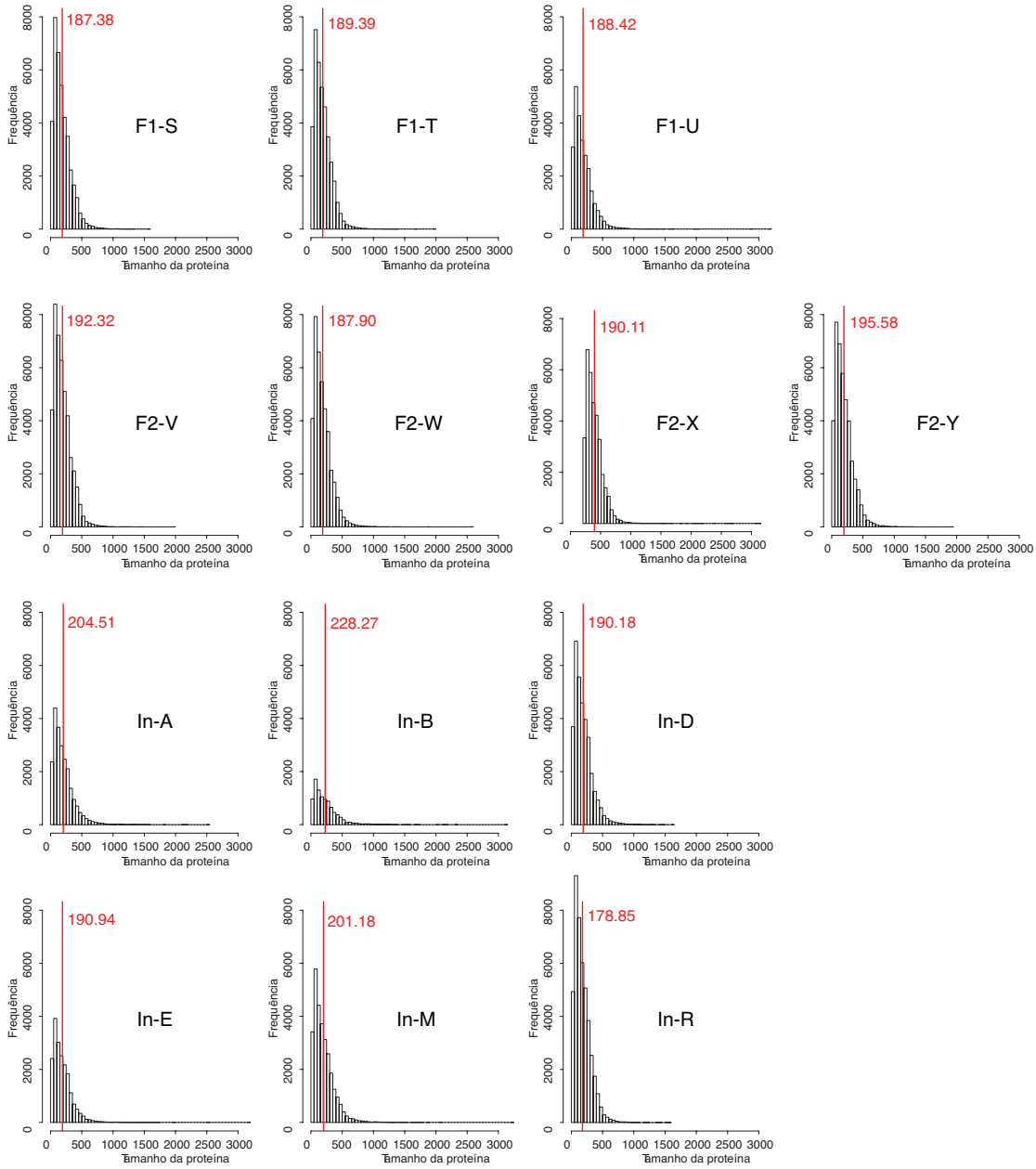

Figura 20 - Histograma mostrando o tamanho das proteínas e suas frequências. A linha vermelha representa a média do tamanho em número de aminoácidos.

Essas proteínas preditas foram então submetidas a uma busca por homologia nas bases de dados KO e UEKO. O percentual de proteínas anotadas varia de cerca de 24% até 36%, com uma considerável colaboração das proteínas enriquecidas pela integração. Esses números sobre proteínas anotadas e grupos de ortólogos associados estão descritos na Tabela 11. Em média, 114 grupos de ortólogos a mais têm associação quando UEKO é utilizado em comparação ao uso de KO.

Tabela 11 - Dados da anotação funcional dos metagenomas usando KO e UEKO como base de dados para busca de homologia.

Nome	Proteínas anotadas no KO	Anotação adicional no UEKO	Percentual de anotação (UEKO)	Grupos do KO com associação	Grupos do UEKO com associação
F1-S	1210			256	265
	1	242	32,00.	1	8
F1-T	1326			284	298
	1	241	35,62.	0	9
F1-U				339	352
	8253	96	32,57.	0	3
F2-V	1582			310	324
	7	259	36,80.	5	8
F2-W	1247			264	275
	5	231	32,79.	2	4
F2-X	1103			259	271
	8	216	33,09.	2	1
F2-Y	1304			261	271
	1	236	32,44.	8	1
In-A				211	221
	6146	159	27,86.	7	2
In-B				146	153
	2216	35	24,19.	6	8
In-D	1046			237	249
	1	245	31,62.	9	0
In-E				212	223
	6113	105	32,29.	7	1
In-M				364	377
	8634	177	30,28.	8	2
In-R	1436			268	281
	0	310	33,58.	2	8

A análise comparativa da anotação usando KO e UEKO como referência, tem diversos pontos a serem analisados. Um deles é a melhoria do escore do BLAST quando usamos uma base de dados com mais informações. Das 219.018 proteínas anotadas por ambas as bases 56,8% delas tiveram uma melhoria no escore do BLAST. Comparando os escores quando utilizamos três bases de dados - KO, UEKO e NR - observamos que mais da metade das entradas do UEKO condizem com o melhor escore da base NR, porém ainda existem muitas entradas a serem adicionadas a base de dados, uma vez que cerca de 32% das anotações com o NR tem uma melhoria no escore de mais de 50% quando comparado com o UEKO. Esses valores para o indivíduo F1-S pode ser visto na Tabela 12. O perfil da anotação para cada sequência original das amostras pode ser visto na Figura 21. Uma pequena parcela (em laranja) tem aumento bastante significativo.

Tabela 12 - Porcentagem de proteínas dentro da faixa de melhoria do escore quando comparadas duas bases de dados.

Melhoria no escore (%)	Bases de dados comparadas	
	KO e UEKO	UEKO e NR
Mesmo escore	38,93%	51,85%
até 10%	27,53%	5,91%
10% a 20%	10,98%	2,76%
20% a 30%	6,48%	2,55%
30% a 40%	4,00%	2,12%
40% a 50%	3,31%	2,00%
mais que 50%	8,77%	32,81%

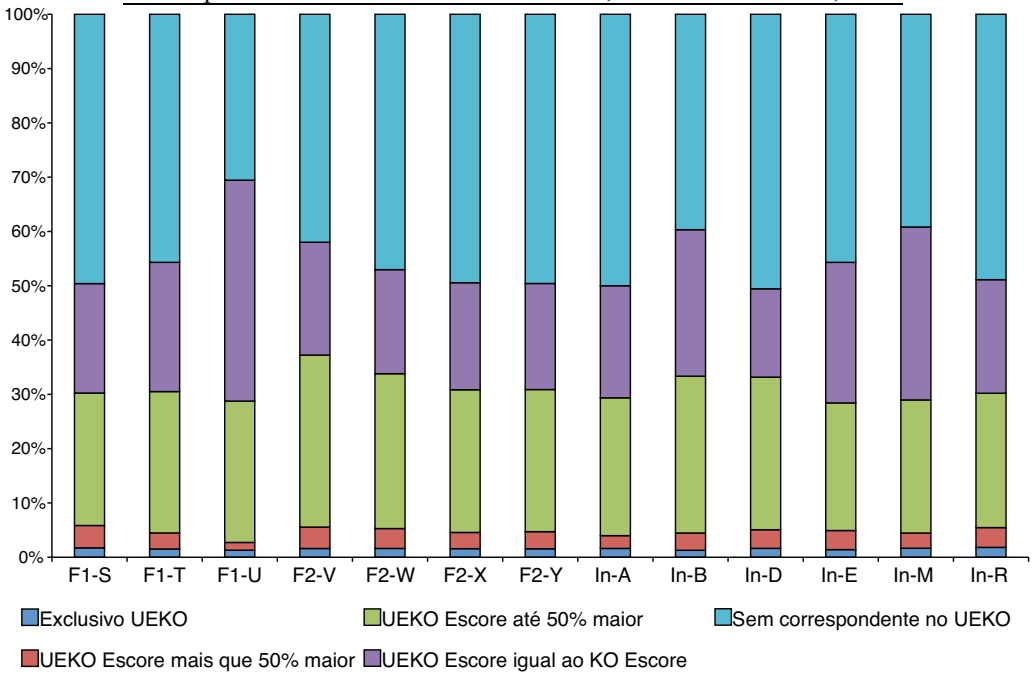

Figura 21 - Perfil da anotação das sequências das amostras metagenômicas. O total de sequências geradas foi classificado quanto a sua anotação, mostrando sequências que não foram anotadas, a comparação do escore entre KO e UEKO, assim como as sequências anotadas exclusivamente no UEKO.

Diante de pequenas variações quantitativas entre as anotações com KO e UEKO, observamos que uma mudança mais sensível é quanto à qualidade da informação agregada. Uma análise dos componentes principais, mostrada na Figura 22, indica que a disposição dos indivíduos no espaço é bastante semelhante ao mudarmos a base de dados, entretanto com o UEKO temos uma melhor distinção dos pontos. Mas o perfil geral é mantido com um eixo orientando os indivíduos infantes (In-E, In-B, In-M e F1-U). Observamos que alguns pontos representando os grupos de ortólogos ganham mais importância como guia para alguns indivíduos ao usarmos o UEKO, como K00936. Da mesma forma, alguns agrupamentos perdem essa importância com o enriquecimento da base, é o caso dos grupos K06147 e K02004.

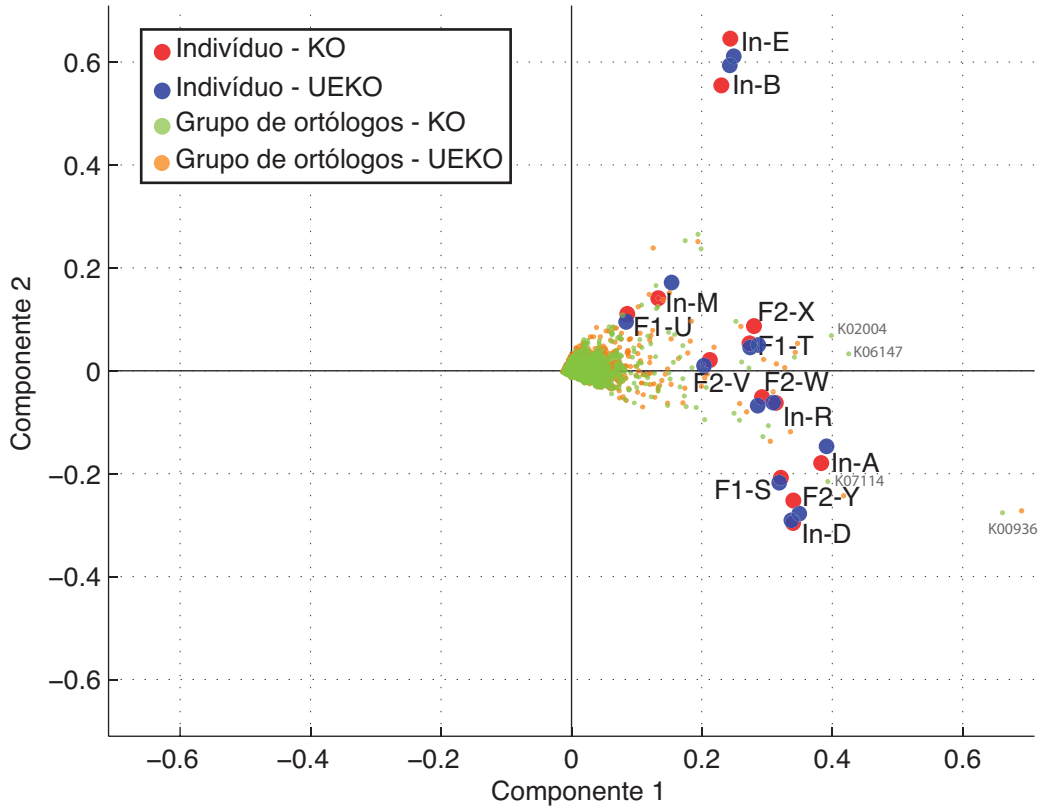

Figura 22 - Análise dos componentes principais das amostras metagenômicas em nível funcional.

Analisando as categorias funcionais das hierarquias do KEGG, identificamos 23 com significativa representatividade (acima de 1% de ocorrência). Funções como Metabolismo de Aminoácidos e Metabolismo de Carboidratos tem uma alta contribuição para o aparato funcional de todos os indivíduos e não apresentam grandes variações quando analisamos com bases de dados diferentes.

A categoria de Metabolismo e Biodegradação de Xenobióticos apresentou um aumento na sua representação ao ser anotada pelo UEKO, enquanto o Metabolismo de Outros Aminoácidos diminuiu sua contribuição total. Entretanto tais mudanças não são estatisticamente significativas. A distribuição para essas 23 categorias pode ser vista na Figura 23.

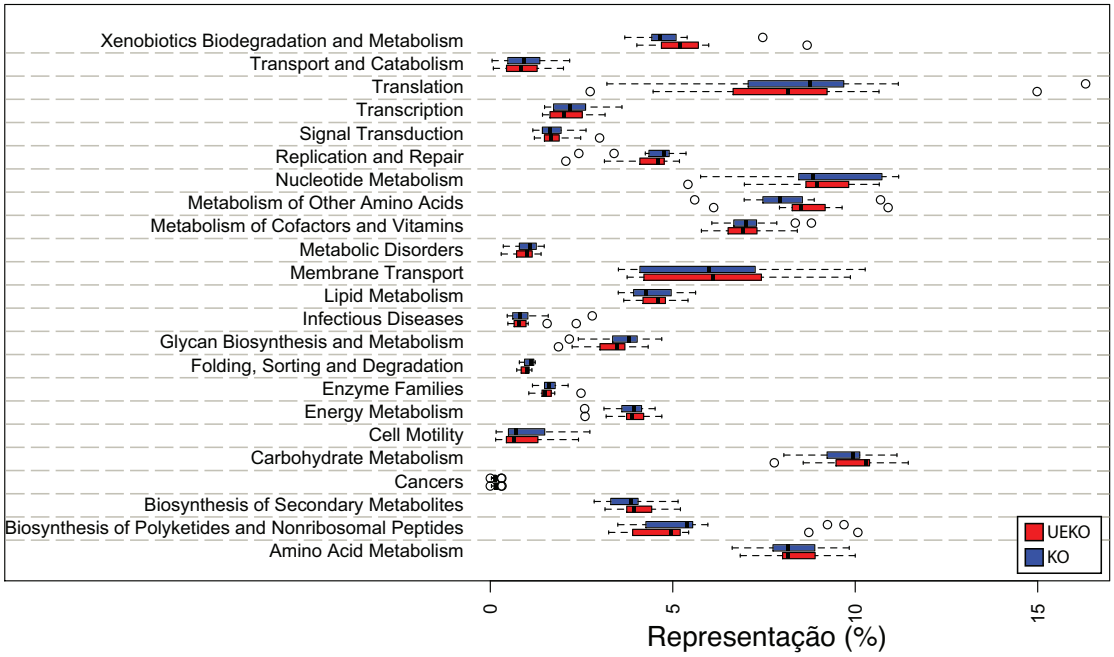

Figura 23 - Gráfico de caixas mostrando contribuição de cada categoria funcional do KEGG para o aparato metabólico das amostras. Os nomes das categorias foram mantidos em inglês para que a consulta ao KEGG seja precisa.

Comparando as proteínas com anotação em ambas as bases de dados observamos que 2.624 delas (1,2%) trocaram de agrupamento de ortólogo. Dessas, 976 (37% das trocas) mudaram também para um grupo de categoria funcional diferente da anotada pelo KO.

Mantendo a mesma categoria, destacamos uma alta freqüência nas trocas internas aos grupos de Transdução de Sinal, Transporte de Membrana, e Metabolismo de Carboidratos. Uma dessas trocas intra-categoria freqüentes acontece no Transporte de Membrana, entre os grupos: K06147, um cassete de ligação de ATP da subfamília B anotado pelo KO e K06148, um cassete também, mas da subfamília B, este anotado pelo UEKO. Essa troca aconteceu 33 vezes e tem um respaldo de uma melhoria do escore do BLAST em média de 27% no UEKO.

Uma visualização das trocas em um nível hierárquico de categoria funcional pode ser visto na Figura 24.

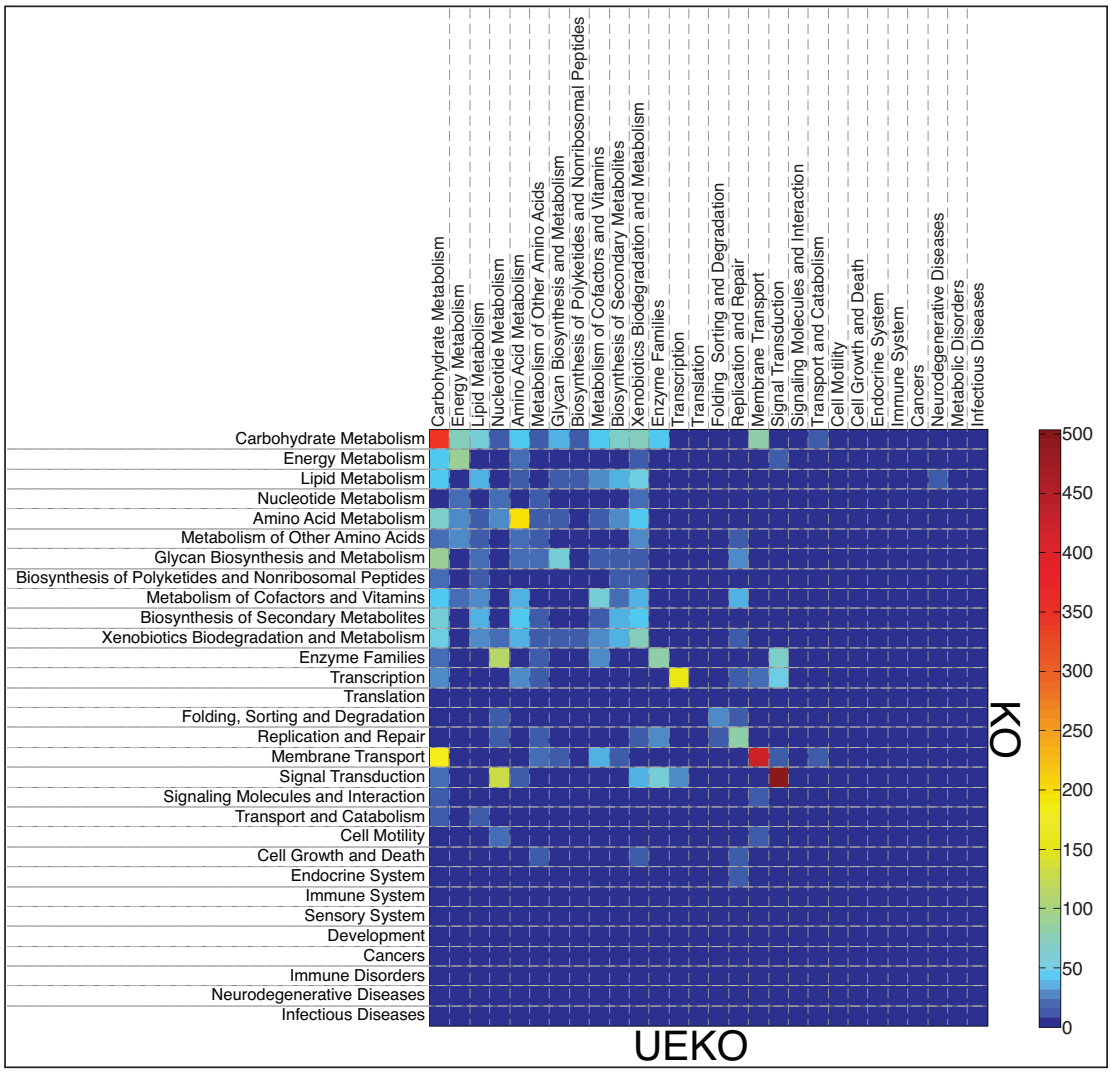

Figura 24 - Visualização das associações de grupo de ortólogo alocadas em categorias funcionais. No eixo horizontal temos a categoria funcional do KO ao qual uma sequencia foi anotada. No eixo vertical está a categoria funcional do grupo para o qual a sequência foi anotada ao usarmos o UEKO como base de dados. Os pontos coloridos mostram a freqüência de ocorrências em que a associação aconteceu (mensurada pela escala de cores à direita). As trocas para grupos de ortólogos dentro da mesma categoria funcional são mostradas na diagonal.

Cada uma dessas trocas analisadas por categoria é então analisada no seu nível mais basal: o agrupamento KO. Onde o universo dessas trocas está concentrado pode ser visto na Figura 25. Uma troca bem freqüente acontece entre os grupos K04763 (integrase/recombinase XerD) e K03733 (integrase/recombinase XerC). Um total de 39 proteínas metagenômicas foram associadas ao XerC pelo KO e ao XerD pelo UEKO, e o escore dessa anotação aumentou em média 3,6%. O contrário também acontece 16 vezes, com o KO anotando para a proteína XerD e o UEKO para XerC, com um aumento médio de 2,8% no escore. Isso mostra que apesar da grande semelhança entre as informações desses agrupamentos o UEKO

acrescenta informações capazes de auxiliar na distinção entre um grupo e outro.

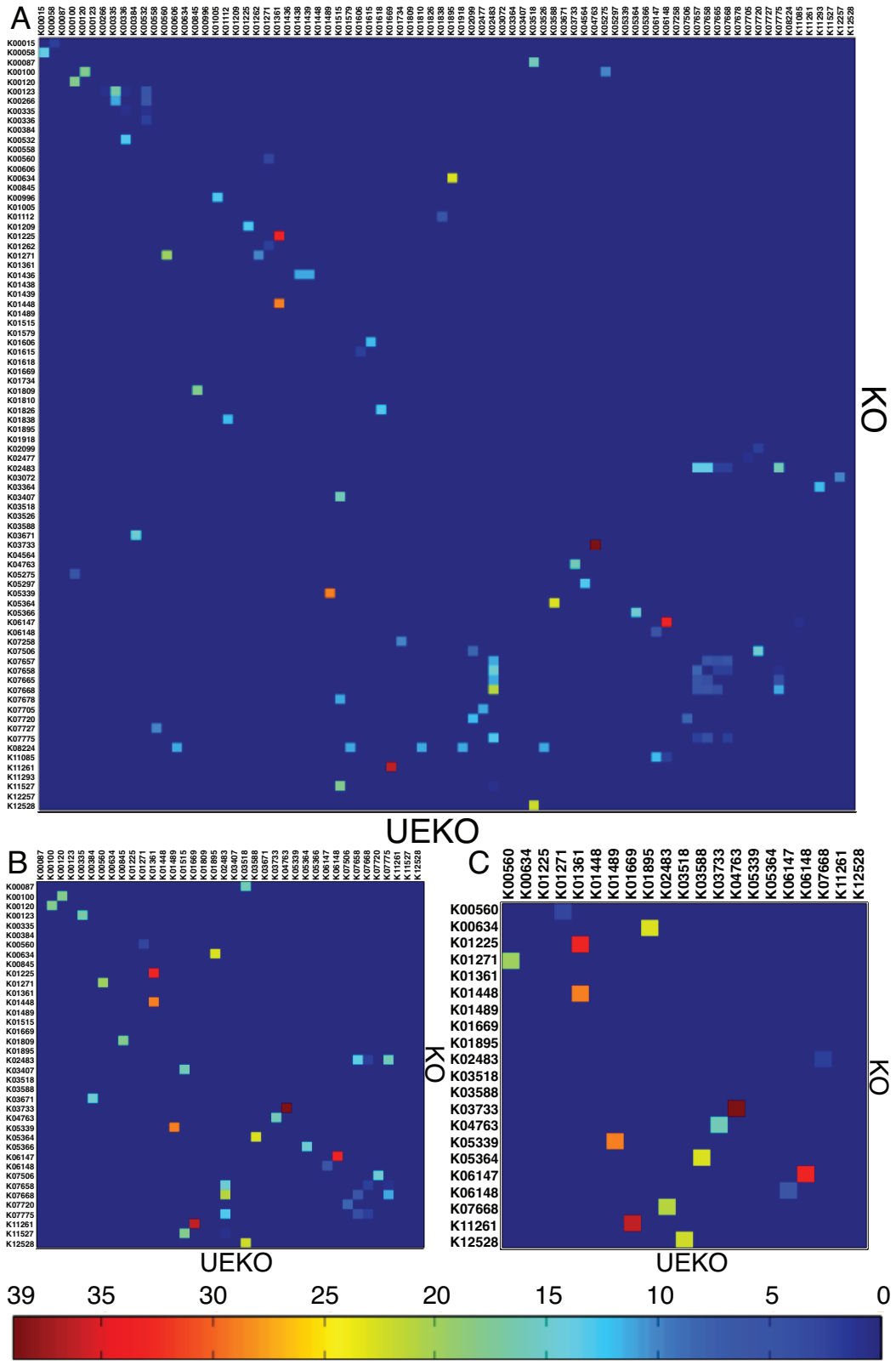

Figura 25 - Visualização das trocas entre grupos de ortólogos. Em A temos todas as trocas, em B as trocas com mais que 10 ocorrências, e em C damos ênfase às trocas com mais que 15 ocorrências.

Outra considerável troca de grupos acontece quando o KO anota a proteína no grupo K00532 (Ferredoxina hidrogenase, EC:1.12.7.2) e o UEKO anota no grupo K00336 (nuoG, subunidade G da NADH desidrogenase I, EC:1.6.5.3). Esse disputa aconteceu 13 vezes e em todas elas a entrada B0MAD8 do UniProtKB é a responsável pela anotação pelo UEKO. Análise manual mostra que essa proteína alinha com [Fe] hidrogenases e Rubrerythrin, e a troca acontece por uma opção na produção do KO que agrupa [Fe] hidrogenases no K00336, como pode ser visto na árvore filogenética mostrada na Figura 26 (ramo marcado em vermelho). Com isso, o processo de enriquecimento recruta [Fe] hidrogenases para o grupo K00336, incluindo a entrada B0MAD8, que fornece melhor alinhamento com as proteínas de metagenômica. Curiosamente, o COG contém um grupo denominado Rubrerythrin (COG1592) e a B0MAD8 está recrutada nele. Em resumo, uma edição do KO posicionando as [Fe] hidrogenases em um grupo específico Rubrerythrin forneceria uma anotação mais apropriada, todavia o UEKO fornece uma anotação ao um ramo da árvore do K00336 que realmente agrupa [Fe] hidrogenases.

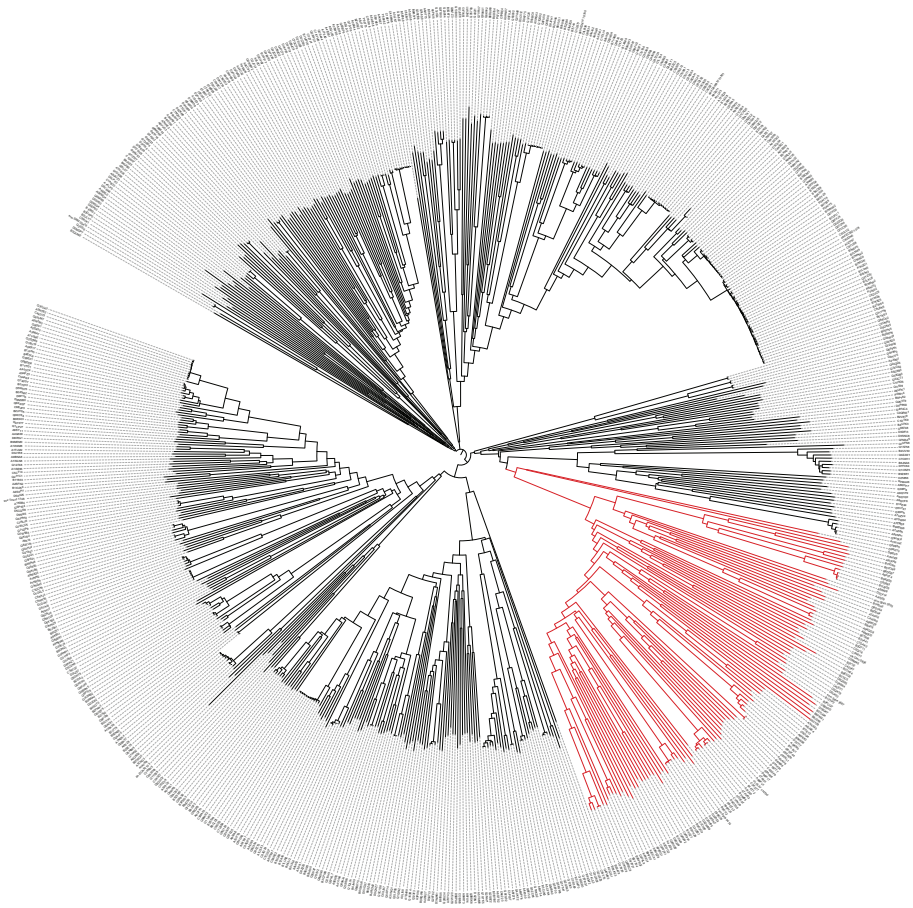

Figura 26 - Árvore filogenética do grupo K00336 mostrando a presença de [Fe] hidrogenases em sua composição original. Em vermelho estão marcadas as ferredoxina-hidrogenases que estão incorretamente classificadas nesse grupo, denominado subunidade G da NADH desidrogenase I.

Entretanto muitas dessas trocas estão associadas à melhoria da anotação proporcionada pelo UEKO. Como exemplo temos uma metaproteína que anotada pelo KO ao grupo K01844, denominado beta-lysine 5,6-aminomutase. O alinhamento é com a entrada UniProt Q21RL6, uma beta-lisina 5-6 aminomutase, produzida por uma *Rhodoferax ferrireducens* que é uma bactéria fototrófica (Bryant e Frigaard, 2006) e por isso incapaz de colonizar o intestino humano. Quando utilizamos o UEKO como referência para a anotação tivemos essa mesma metaproteína anotada pela entrada C7IVJ3 associada ao grupo K01846 do UEKO, quem contém metilaspartato mutases. O escore da anotação aumentou em 4,6 vezes, indo de 123 atribuído pelo KO para 570 bits atribuídos pelo UEKO. Essa proximidade maior do grupo K01846 pode ser vista na Figura 27.

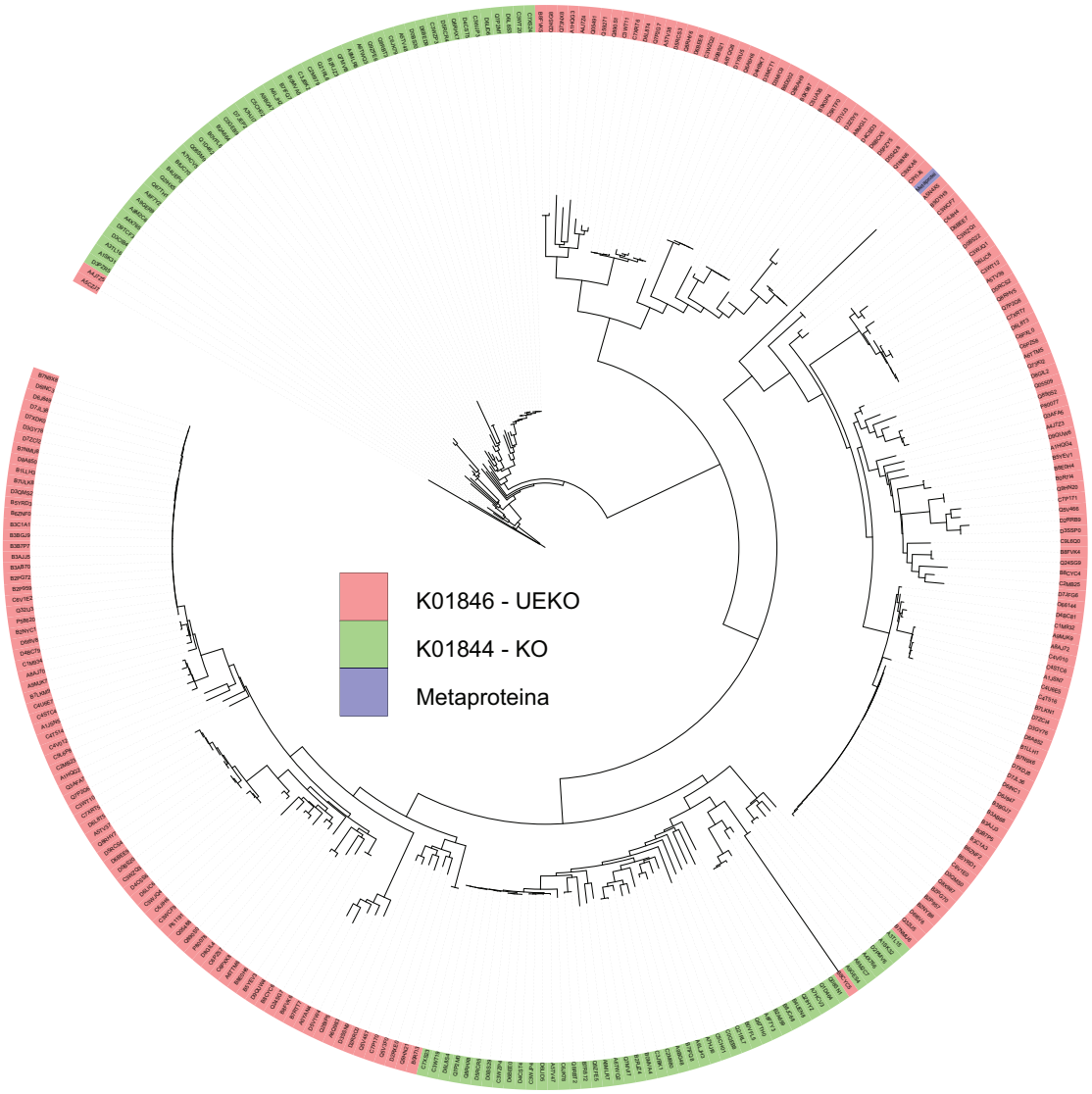

Figura 27 - Cladograma evidenciando a semelhança da metaproteína ao grupo K01846. A metaproteína (azul) encontra-se mais associada com o grupo anotado pelo UEKO.

A proteína usada pelo UEKO para anotar a metaproteína é de uma *Thermoanaerobacter ethanolicus*, uma bactéria da classe Clostridia, associada ao gênero *Clostridium*, bastante comum na microbiota. Essa observação da proximidade dos clados nos levou a analisar o ancestral comum que une os organismo anotados pelo KO e pelo UEKO. Nesse nosso exemplo o ancestral comum pertence à categoria de super reino Bacteria, uma vez que o KO anotou para o filo Proteobacteria e o UEKO para o filo Firmicutes. Foram então comparados todas as associações filogenéticas entre os anotadores KO e UEKO para observarmos o ancestral comum, e ver quão distante filogenéticamente vai a influência enriquecimento da base. Na Figura 28 (A) temos o número de ocorrências para cada nível cladístico quando comparamos os organismos em que a anotação KO e UEKO não coincidiram, ou seja, os grupos foram trocados. Em (B) temos a mesma análise, mas com os organismos em que a anotação pelas duas bases de dados coincidiu. Surpreendentemente, mesmo quando a anotação é a mesma, o UEKO fornece proteínas para análise comparativa tão distantes quanto de reinos diferentes (agrupadas pelo mesmo super reino).

Com a anotação funcional a grupos de ortólogos, foi possível a identificação de um genoma mínimo, ou seja, genes da microbiota que estavam presentes em todos os indivíduos, o que pode ser entendido como as funções mínimas exigidas para a sobrevivência do indivíduo e da sua microbiota. Ao utilizarmos o KO para a anotação identificamos 860 grupos de ortólogos presentes em todos os 13 indivíduos, e esse número aumenta para 899 quando usamos o UEKO como base de dados. Essas funções que foram acrescentadas pelo enriquecimento da base de dados podem ser vistas na Figura 29, onde se destaca o papel do UEKO ao acrescentar informações de Transportadores ABC, Sistemas de dois componentes, além de preencher lacunas ligando com isso processos até então separados fisicamente no mapa metabólico.

Devido à grande diferença filogenética e funcional entre os quatro infantes (nos quais a microbiota estaria em formação) e os demais indivíduos (quatro crianças e cinco adultos) como mostrado nas Figuras 19 e 22, a análise do genoma mínimo foi feita em dois grupos, infantes e não infantes. As funções sempre presentes nos não infantes são 1.447 quando usamos KO e 1.502 ao usarmos UEKO para anotação. Já para os infantes o número de agrupamentos é de 1.102 com KO e 1.152 usando o UEKO. Identificamos 603 funções presentes no genoma mínimo dos não infantes, mas ausentes no de infantes; e 253 funções do genoma mínimo de infantes ausentes no de não infantes.

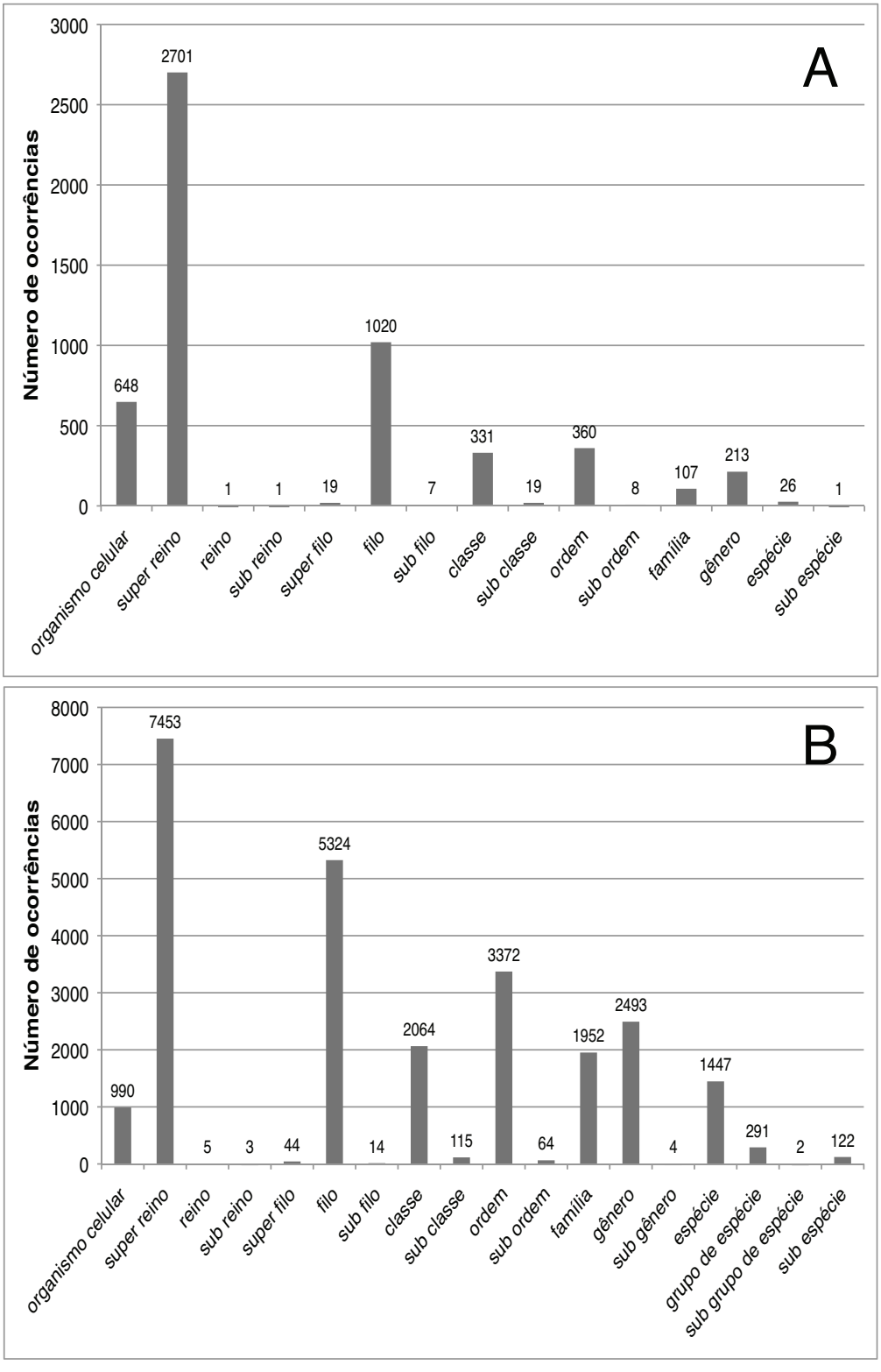

Figura 28 - Distribuição da ancestralidade comum entre proteínas do KO e UEKO usadas para a anotação metagenômica. Em A temos, para cada categoria cladística, o número de associações entre organismo anotador KO e UEKO para os casos em que o grupo de ortólogos mudou quando mudamos a base de dados. Em B temos a mesma análise de ocorrências de ancestralidade comum entre anotadores de KO e UEKO apontando para uma categoria cladística, porém consideramos as relações em que não houve troca de grupo de ortólogs.

Dentre as funções exclusivas ou mesmo preferenciais dos infantes (presente em todos

infantes e no máximo em um não infante) identificamos várias proteínas de transporte e de membrana, mas duas em especial: K01920 (glutationa sintase) presente em todos infantes e em apenas um não infante (F1-T, uma mulher de 28 anos); e K11748 (sistema de efluxo de potássio regulado por glutationa) que está ausente em todos os não infantes. Isso sugere a necessidade de interação com o meio ambiente nas etapas de estabilização da microbiota em seu hospedeiro, além da sustentabilidade do sistema, uma vez que é a mesma é capaz de produzir um componente regulador do seu sistema de efluxo.

Outro aspecto importante observado nessas amostras é a complementaridade entre os genomas da microbiota e o humano. Analisamos as reações realizadas por cada grupo de ortólogo localizando-as no mapa metabólico. Mapeamos grupos identificados com UEKO para o genoma mínimo da microbiota e para o ser humano. No mapa metabólico, algumas vezes uma reação é possível de ser realizada por mais de um grupo KO, portanto há redundância. Fazendo esse mapeamento para microbiota e humanos, identificamos 29 compostos que atuavam entre os genomas, ou seja, utilizam produtos exclusivos de reações da microbiota como único substrato possível para as reações humanas. Dentre esses produtos da microbiota identificamos 3 vitaminas (Biotina, Pantotenato e Riboflavina) e vários outros como D-Glucosamina, Tiamina monofosfato, Diidropteroato, etc. Esse mutualismo é muito importante para a produção de vitaminas, aminoácidos e para a degradação de xenobióticos, como pode ser visto na Figura 30.

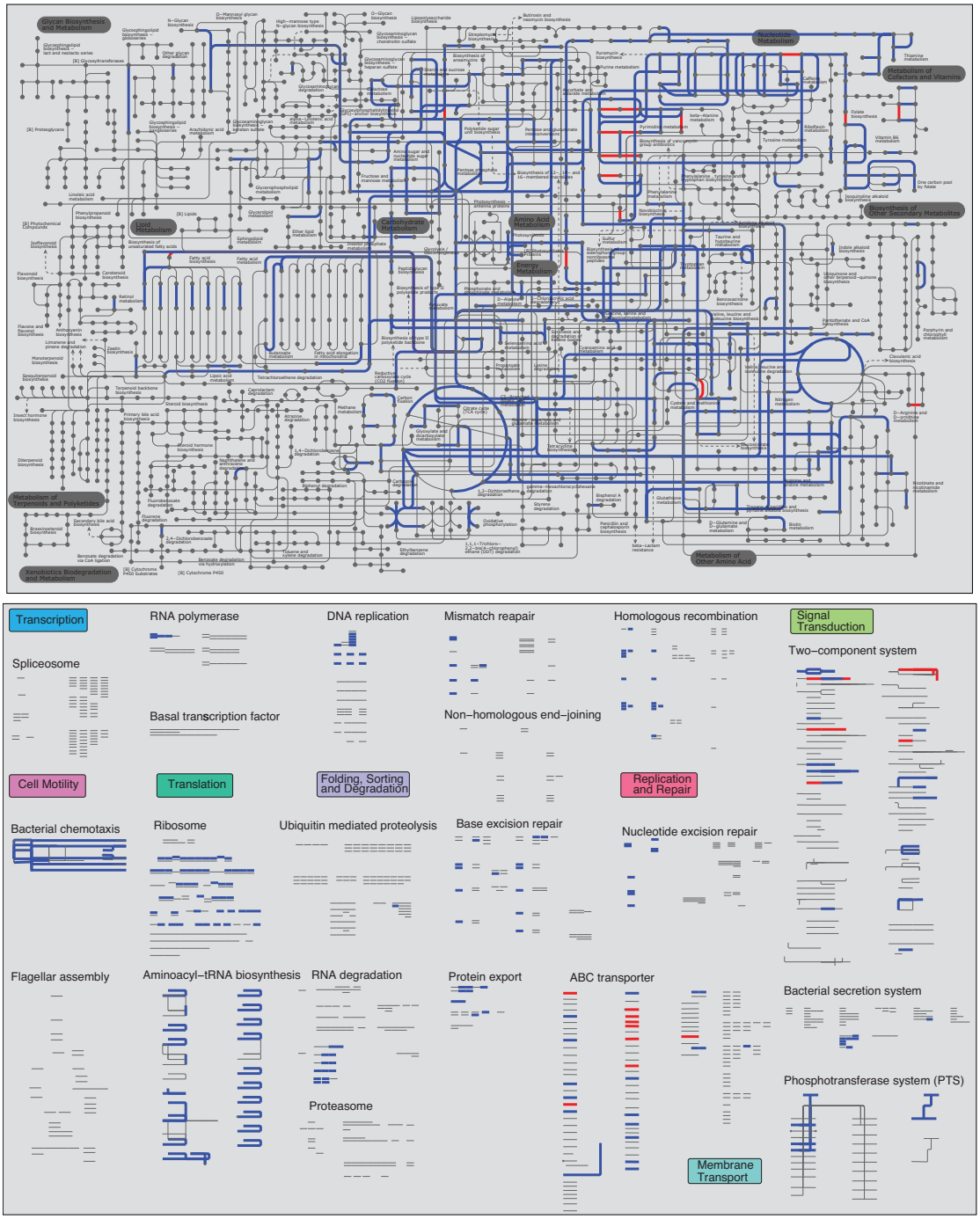

Figura 29 - Mapeamento funcional dos produtos do genoma mínimo da microbiota humana identificados por KO (azul) ou exclusivamente pelo UEKO (vermelho).

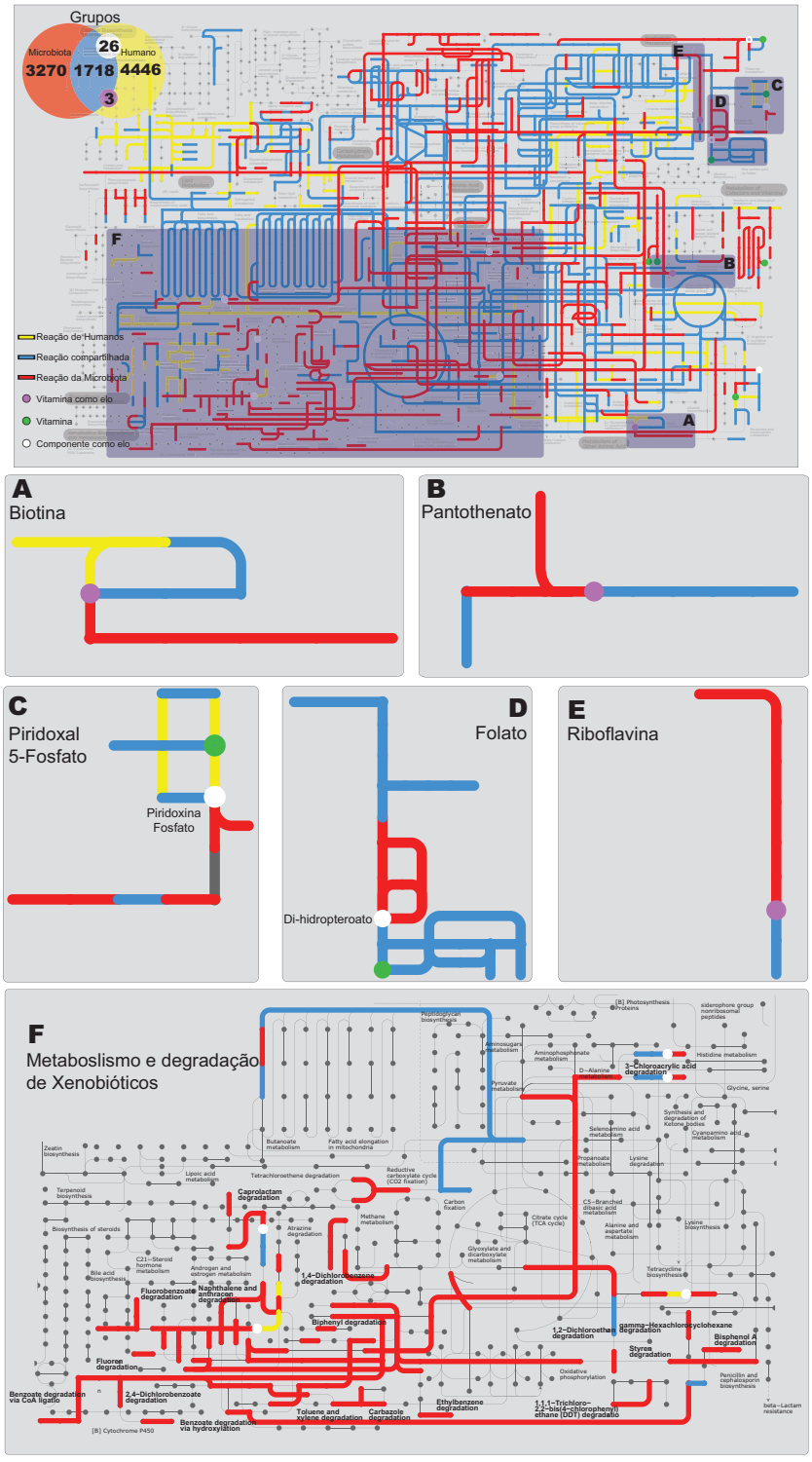

Figura 30 - Mapa metabólico mostrando a complementaridade entre o genoma humano e sua microbiota. Em destaque temos as reações que levam a produção de vitaminas e da degradação de xenobióticos, vias nas quais encontramos mais elos que ligam o nosso genoma à microbiota. Compostos marcados em vermelho (exclusivos da microbiota) geram produtos (círculos) que são substratos para compostos exclusivos do mapa humano (em amarelo) ou presentes também na microbiota (em azul). Figura superior: mapa metabólico. Painéis A-F: detalhe das vias indicadas. No painel superior estão indicadas as quantidades de grupos KO utilizados no mapeamento dos compostos do mapa. Números destacados: 29 produtos exclusivos da microbiota, sendo três vitaminas, são conectados a composto do mapa humano.

4.7 OS ENTEROTIPOS DA MICROBIOTA INTESTINAL

Sob posse das sequências geradas pelo consórcio MetaHIT foram feitas análises filogenéticas e funcionais das amostras de microbiota. Um total de 39 indivíduos foram amostrados, dentre eles os 13 japoneses estudados anteriormente. As informações como idade, sexo, nacionalidade e estado clínico de cada uma das novas amostras estão na Tabela 13.

Tabela 13 - Detalhes sobre os indivíduos estudados.

ID da amostra	Projeto	Nome da amostra	Nacionalidade	Sexo	Idade	Estado Clínico
DA-AD-1	MetaHIT	MH6	dinamarquês	F	59	saudável
DA-AD-2	MetaHIT	MH13	dinamarquês	M	54	saudável
DA-AD-3	MetaHIT	MH12	dinamarquês	F	49	obeso
DA-AD-4	MetaHIT	MH30	dinamarquês	M	59	obeso
ES-AD-1	MetaHIT	CD1	espanhol	F	25	IBD
ES-AD-2	MetaHIT	CD2	espanhol	M	49	saudável
ES-AD-3	MetaHIT	UC4	espanhol	F	47	IBD
ES-AD-4	MetaHIT	UC6	espanhol	F	38	saudável
FR-AD-1	MicroObes	NO1	francês	M	63	saudável
FR-AD-2	MicroObes	NO3	francês	M	61	saudável
FR-AD-3	MicroObes	NO4	francês	M	60	saudável
FR-AD-4	MicroObes	NO8	francês	M	60	saudável
FR-AD-5	MicroObes	OB2	francês	M	64	obeso
FR-AD-6	MicroObes	OB1	francês	M	63	obeso
FR-AD-7	MicroObes	OB6	francês	M	62	obeso
FR-AD-8	MicroObes	OB8	francês	M	60	obeso
IT-AD-1	MicroAge	A	italiano	F	84	idoso
IT-AD-2	MicroAge	B	italiano	M	87	idoso
IT-AD-3	MicroAge	C	italiano	F	77	idoso
IT-AD-4	MicroAge	D	italiano	M	80	idoso
IT-AD-5	MicroAge	E	italiano	M	70	idoso
IT-AD-6	MicroAge	G	italiano	F	72	idoso
JP-AD-1	kurokawa07	F1-S	japonês	M	30	saudável
JP-AD-2	kurokawa07	F1-T	japonês	F	28	saudável
JP-AD-3	kurokawa07	F2-V	japonês	M	37	saudável
JP-AD-4	kurokawa07	F2-W	japonês	F	36	saudável
JP-AD-5	kurokawa07	F2-X	japonês	M	3	saudável
JP-AD-6	kurokawa07	F2-Y	japonês	F	1,5	saudável
JP-AD-7	kurokawa07	In-A	japonês	M	45	saudável
JP-AD-8	kurokawa07	In-D	japonês	M	35	saudável
JP-AD-9	kurokawa07	In-R	japonês	F	24	saudável
JP-IN-1	kurokawa07	F1-U	japonês	F	0,58	saudável
JP-IN-2	kurokawa07	In-B	japonês	M	0,5	saudável
JP-IN-3	kurokawa07	In-E	japonês	M	0,25	saudável
JP-IN-4	kurokawa07	In-M	japonês	F	0,33	saudável
AM-AD-1	gill06	Subject7	americano	F	28	saudável
AM-AD-2	gill06	Subject8	americano	M	37	saudável
AM-F10-T1	turnbaugh09	F10T1Ob1	americano	F		obeso
AM-F10-T2	turnbaugh09	F10T2Ob1	americano	F		obeso

Todas essas sequências foram submetidas a uma montagem utilizando o *pipeline* do programa SMASH (Arumugam, Harrington *et al.*, 2010). Os resultados dessa montagem são mostrados na Tabela 14.

Tabela 14 - Resultado da montagem dos 39 metagenomas.

Amostra	Sequências	*Singletons*	*Contigs*	Proteínas
DA-AD-1	237710	85700	19816	105516
DA-AD-2	224711	80256	18910	99166
DA-AD-3	231024	88736	21465	110201
DA-AD-4	227411	91405	22135	113540
ES-AD-1	223746	50190	14898	65088
ES-AD-2	230738	69752	15257	85009
ES-AD-3	236855	78396	20260	98656
ES-AD-4	229783	90695	24863	115558
FR-AD-1	125260	66486	15390	81876
FR-AD-2	113507	61151	12439	73590
FR-AD-3	115862	55637	14694	70331
FR-AD-4	120268	72738	14808	87546
FR-AD-5	129745	70637	13294	83931
FR-AD-6	118423	64043	14112	78155
FR-AD-7	118172	56166	14994	71160
FR-AD-8	112592	64959	12266	77225
IT-AD-1	116244	43644	13489	57133
IT-AD-2	115636	47103	12461	59564
IT-AD-3	116746	57795	16029	73824
IT-AD-4	116891	31691	6606	38297
IT-AD-5	118227	62846	14236	77082
IT-AD-6	116085	61669	13766	75435
JP-AD-1	78123	16561	14535	31096
JP-AD-2	80477	22788	14961	37749
JP-AD-3	79846	20442	17351	37793
JP-AD-4	78670	17634	13537	31171
JP-AD-5	79773	19383	12302	31685
JP-AD-6	79357	21669	15134	36803
JP-AD-7	75532	15765	5327	21092
JP-AD-8	80627	28252	10390	38642
JP-AD-9	81346	17969	16420	34389
JP-IN-1	80796	11452	6136	17588
JP-IN-2	79972	5120	1671	6791
JP-IN-3	79787	10324	5647	15971
JP-IN-4	87324	11137	6665	17802
AM-AD-1	65042	34718	7113	41831
AM-AD-2	74452	27947	9501	37448
AM-F10-T1	248939	117041	33379	150420
AM-F10-T2	435911	132093	46287	178380

Essas sequências foram submetidas a assinatura utilizando a base nt adicionada de 1551 genomas referência (genomas completos comumente encontrados na microbiota) para a sua classificação taxonômica. Essa caracterização mostra que realmente *Firmicutes* e

Bacteroidetes são os filos que constituem ampla maioria da microbiota humana, com o gênero *Bacteroides* sendo o mais abundante a ao mesmo tempo o mais variável ao longo das amostras. Cerca de 52,8% das sequências foram associadas a um gênero, e a distribuição dos principais gêneros e filos pode ser vista na Figura 31.

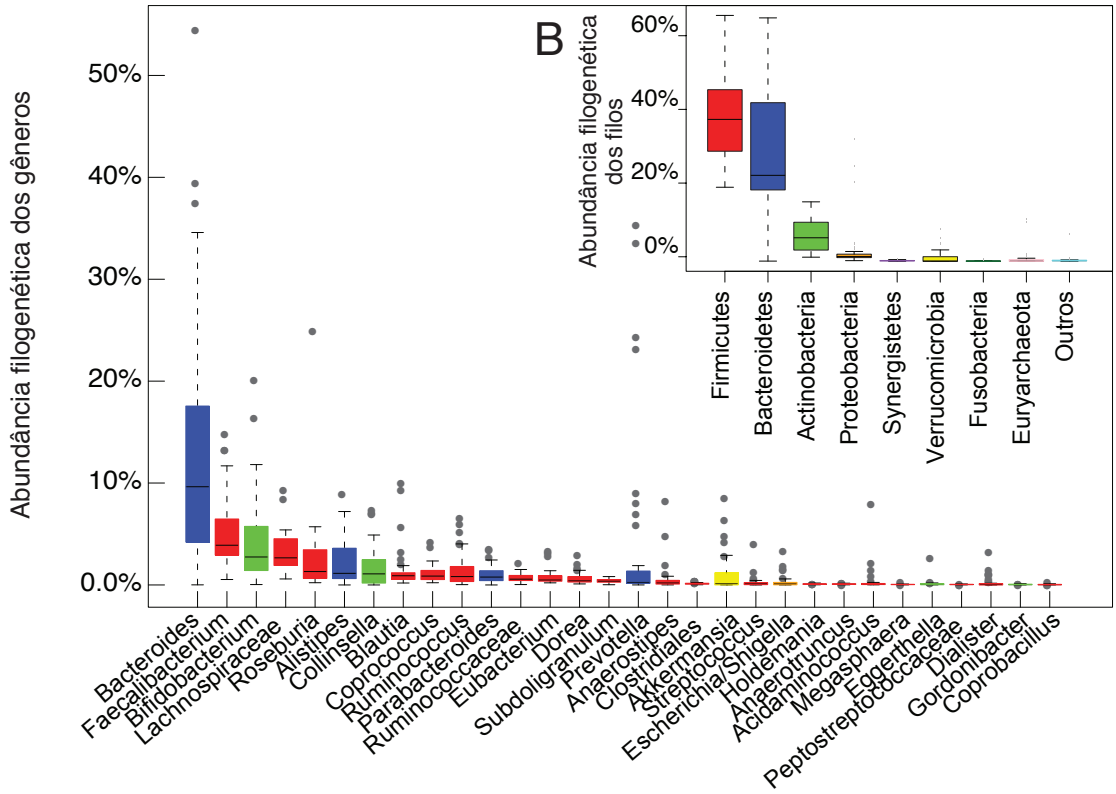

Figura 31 - Gráfico de caixa mostrando a contribuição dos 30 gêneros mais abundantes e seus filos. As cores das barras condizem com o filo a que cada gênero pertence como indicado no gráfico em B.

A anotação funcional utilizou o protocolo SMASH e duas principais bases de dados como referência: eggNOG e KO. Com isso, 2.279.675 das 3.721.987 proteínas preditas foram anotadas, um percentual de 61,2%.

As bactérias da microbiota vivem sob constante pressão seletiva do hospedeiro e também de outros micróbios. Isso frequentemente leva à homeostase do sistema em que poucas bactérias são abundantes e muitas delas tem uma pequena representatividade diante do montante de organismos vivendo naquele ambiente. As funções mais abundantes nas amostras metagenômicas estão geralmente ligadas a organismos com uma alta representatividade. Porém identificamos algumas atividades associadas preferencialmente a gêneros ou grupos de gêneros (agrupados para facilitar a visualização) que são encontrados em baixa frequencia, constituindo uma pequena porção da microbiota. Por exemplo, o gênero *Escherichia* não é muito abundante, entretanto ele contribui em cerca de 90% com a produção de duas proteínas abundantes no indivíduo IT-AD-5. As proteínas FimA (K07345) e PapC (K12518) são

responsáveis pela produção do pilus, que auxilia na fixação da bactéria na mucosa e é componente chave para a troca de plasmídios pela conjugação, que leva à troca de material genético que pode conferir proteção como por resistência a antibióticos (Krogfelt, 1991). Outro exemplo é uma proteína induzida por danos no DNA, presente em vários indivíduos e produzida principalmente por *Eubacterium* e *Blautia*. As funções, gêneros e contribuições estão na Tabela 15. Esses dados ilustram como uma função pode ser apresentada por organismos pouco representativos, mas podem conter funções suficientemente importantes para serem mantidos.

Tabela 15 - Funções abundantes executadas por gêneros pouco representados.

KO	Descrição	Amostra	Gênero ou grupo de gêneros	Abundância do(s) gênero(s)	Contribuição do(s) gênero(s) para a função
K07473	DNA-damage-inducible protein J	AM-F10-T1	*Eubacterium*	0,74%	55,27%
			Blautia	5,83%	10,24%
			Lachnospiraceae	5,02%	5,27%
			Roseburia	5,21%	5,27%
			Peptostreptococcaceae	0,08%	4,29%
			Ruminococcus	4,02%	3,65%
			Coprococcus	0,83%	2,63%
			Clostridiales	0,15%	2,58%
			Faecalibacterium	8,19%	2,58%
			Bifidobacterium	5,53%	0,35%
			Megasphaera	0,09%	0,35%
		DA-AD-2	*Eubacterium*	0,73%	47,02%
			Faecalibacterium	0,85%	16,89%
			Blautia	0,56%	10,32%
			Lachnospiraceae	1,33%	6,88%
			Clostridiales	0,13%	1,72%
		ES-AD-4	*Blautia*	0,81%	40,31%
			Faecalibacterium	3,70%	15,62%
			Coprococcus	1,05%	12,34%
			Lachnospiraceae	3,47%	11,91%
			Eubacterium	0,42%	4,11%
			Clostridiales	0,17%	3,29%
			Ruminococcus	0,96%	2,81%
		IT-AD-1	*Eubacterium*	0,90%	29,95%
			Blautia	0,80%	21,24%
			Lachnospiraceae	2,66%	16,60%
			Peptostreptococcaceae	0,08%	11,23%
			Ruminococcus	6,71%	5,34%
			Escherichia/Shigella	1,65%	4,83%
		IT-AD-3	*Blautia*	0,99%	27,02%
			Faecalibacterium	7,41%	17,01%
			Coprococcus	1,67%	15,35%
			Eubacterium	0,47%	11,38%
			Clostridiales	0,11%	5,33%
K12518	P pilus assembly protein, porin PapC	IT-AD-5	*Escherichia/Shigella*	2,03%	92,07%
K07345	P pilus assembly protein, pilin FimA	IT-AD-5	*Escherichia/Shigella*	2,03%	96,25%
		JP-AD-2	*Escherichia/Shigella*	1,71%	87,33%
			Citrobacter	0,11%	5,37%

Para ter uma visão geral das amostras nós avaliamos as similaridades dos perfis filogenéticos, representados pela frequencia de gêneros em cada amostra. Dessa análise foram removidos os indivíduos americanos do projeto "gill06" devido a uma incomum ausência de *Bacteroides* e suspeita de artefatos técnicos (Salonen, Nikkila *et al.*, 2010), e as sequências de infantes devido à diferença do perfil de um adulto. . Uma análise multidimensional de grupos e uma análise de componentes principais mostraram que os 13 indivíduos restantes formam três grupos distintos aos quais chamamos de "enterotipos". Cada um desses três enterotipos são identificados pela variação nos níveis de um de três gêneros: *Bacteroides* (enterotipo 1), *Prevotella* (enterotipo 2) and *Ruminococcus* (enterotipo 3), como mostrados na Figura 24 (A) e (D). A mesma análise foi feita com dois conjuntos de dados metagenômicos publicados, de origem diferentes: 85 amostras de microbiota de dinamarqueses sequenciados com um equipamento Illumina (Qin, Li *et al.*, 2010), e pirosequenciamento da região 16S da microbiota de 154 indivíduos americanos (Turnbaugh, Hamady *et al.*, 2009). Os resultados mostram que o conjuto de dados pode também ser representado melhor por três grupos, como pode ser visto na Figura 32 (B) e (C), respectivamente. Dois desses grupos são também orientados principalmente por *Bacteroides* e *Prevotella*, enquanto o terceiro grupo é guiado pela prevalência de organismos relacionados ao gênero *Blautia* e a um conjunto não classificado de *Lachnospiraceae* nos dados 16S rDNA e Illumina, respectivamente. Uma análise de correlação dos 33 indivíduos mostra que a abundância de cada um dos três gêneros principais está fortemente correlacionada à representatividade de outros gêneros, seja coocorrendo ou inibindo os outros organismos, como pode ser visto na Figura 32 (E). Isso sugere que os enterotipos são na verdade guiados por grupos de espécies que juntas contribuem para uma composição da comunidade favorável para a sobrevivência de todos.

Utilizamos então a mesma métrica de agrupamento, com as mesmas 33 amostras, porém avaliando as similaridades dos perfis funcionais, representados agora pela frequencia de cada grupo de ortólogo, como pode ser visto na Figura 33 (A). Esse agrupamento mostrou um agrupamento das amostras semelhante ao anterior em que usávamos assinatura filogenética. Pequenas diferenças foram encontradas, em que 5 amostras encontram-se em um grupo diferente quando comparado com a Figura 32 (A), indicando que agrupamento por função e composição filogenética apresentam algumas discordâncias, como ocorre com a amostra ES-AD-3 que nos grupos baseados na contribuição dos gêneros pertencia ao enterotipo 2, porém quanto à composição funcional é mais similar aos membros do enterotipo 1.

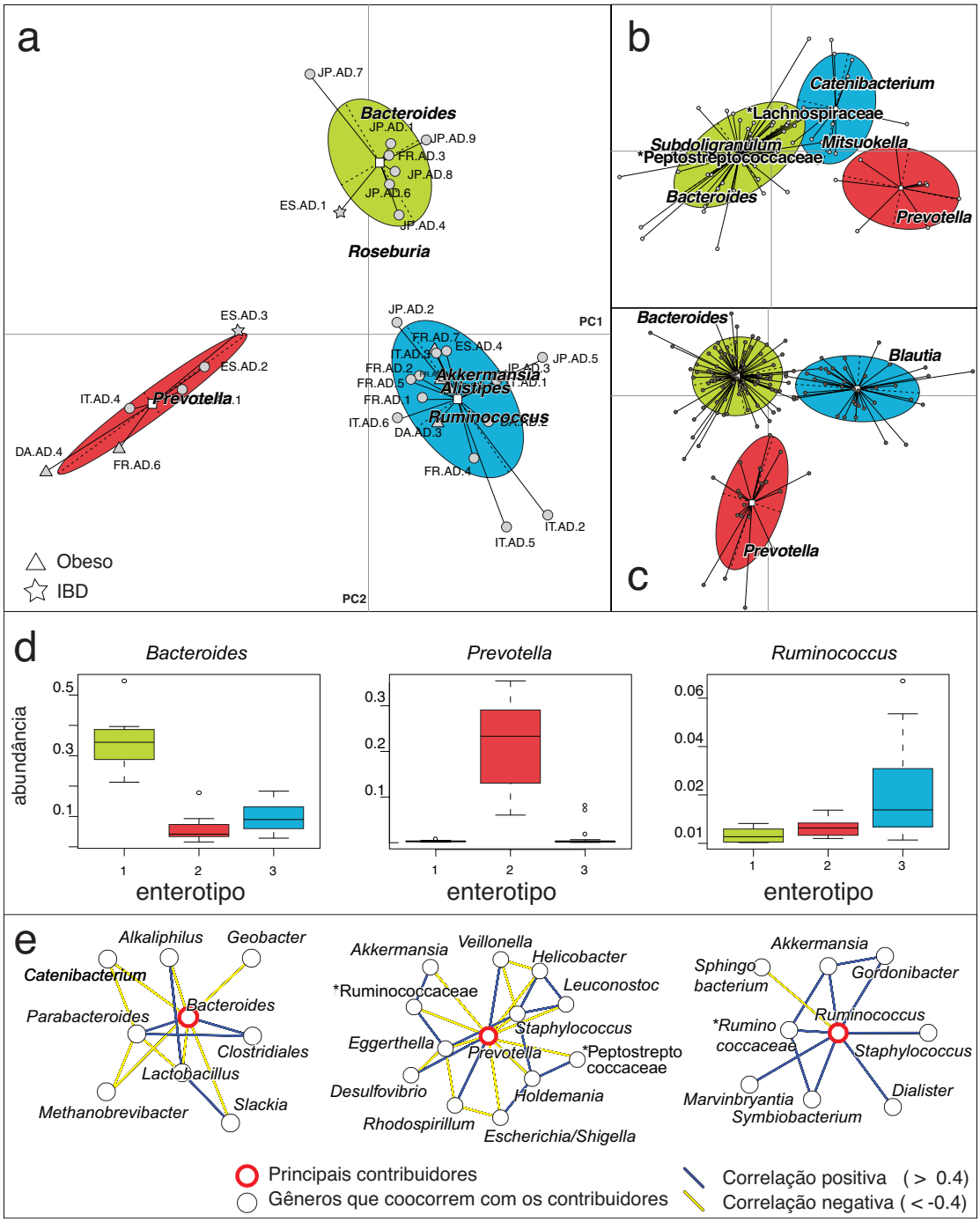

Figura 32 - Visualização e caracterização dos enterotipos de microbiota. (A) 33 amostras de microbiota sequenciadas pelo método de Sanger. (B) Conjunto contendo 85 metagenomas de indivíduos dinamarqueses sequenciados usando Illumina e publicados. (C) Mapeamento da região 16S do metagenoma de 154 indivíduos. A análise do componente principal e a delimitação da elipse é feita pelo pacote "ade4" do programa R e marca com um retângulo o centro de gravidade de cada grupo. (D) Abundância dos principais contribuidores para cada enterotipo identificado nos 33 indivíduos. (E) Rede de coocorrências dos três enterotipos. Gêneros não identificados estão marcados com um asterisco.

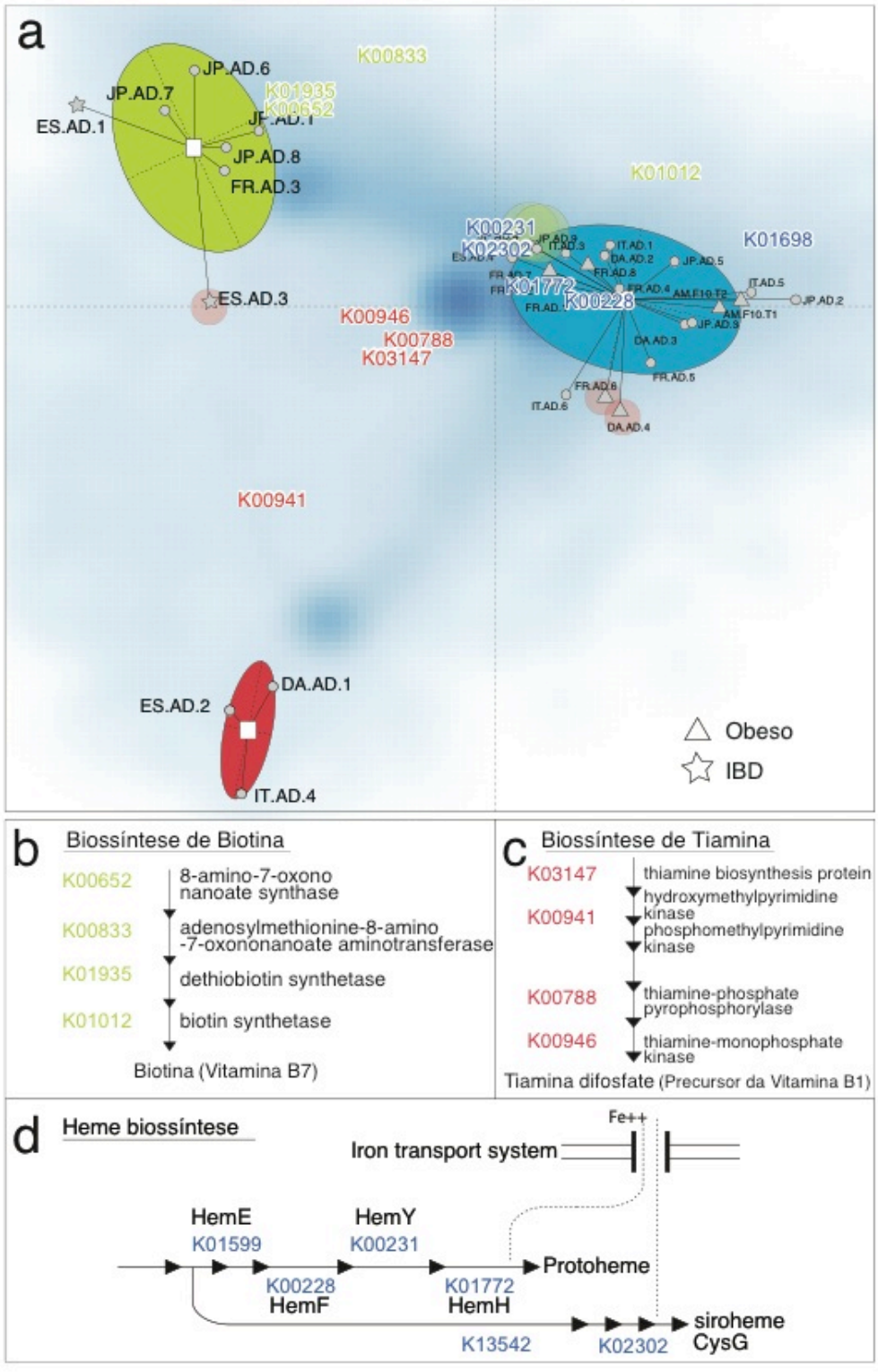

Figura 33 - Agrupamento das amostras metagenômicas baseado em critérios funcionais. (A) Disposição das amostras em torno dos seus grupos utilizando o mesmo pacote "ade4" do programa R utilizado na figura anterior. Os círculos transparentes marcam os indivíduos que mudaram de grupo em relação à análise anterior. A nuvem azul indica uma estimativa da densidade de KO naquelas coordenadas. Os KO selecionados estão destacados. (B) Enzimas da via de biossíntese de biotina que estão supeerrepresentadas no enterotipo 1. (C)

Enzimas da via de biossíntese de tiamina super-representadas no enterotipo 2. (D) Enzimas da biossíntese do grupo Heme super-representadas no enterotipo 3.

Para determinar a base filogenética e funcional dos enterotipos identificados nós investigamos em detalhes as diferenças na composição em níveis de gênero, gene e vias metabólicas, assim como a correlação em abundância dos gêneros que co-ocorrem com os principais.

Assim, o enterotipo 1, contendo 8 indivíduos, é enriquecido com *Bacteroides* ($p<0,01$), genero que co-ocorre, por exemplo, com *Parabacteroides*. O enterotipo 2 consiste em 6 amostras e por sua vez está super-representado com sequências de *Prevotella* ($p<0,01$) co-ocorrendo com *Desulfovibrio*. O terceiro enterotipo é o mais frequente e está enriquecido com *Ruminococcus* ($p<0,01$) tendo a co-ocorrência de *Akkermansia*. As co-ocorrências podem ser vistas na Figura 32, assim como na Tabela 16 com os valores-p para a super-representação.

Tabela 16 - Gêneros super-representados nos enterotipos e os valores-p para seu enriquecimento.

Enterotipo	Gênero	Valor-p
1	*Acidaminococcus*	0
	Bacteroides	**0**
	Roseburia	0
	Faecalibacterium	8,88E-225
	Anaerostipes	2,95E-201
	Parabacteroides	2,70E-150
	Clostridiales	1,27E-18
2	***Prevotella***	**0**
	Streptococcus	2,18E-225
	Enterococcus	2,78E-64
	Desulfovibrio	1,26E-12
	Lachnospiraceae	0,0099863
3	*Akkermansia*	0
	Alistipes	0
	Klebsiella	0
	Ruminococcus	**0**
	Escherichia/Shigella	3,63E-288
	Dialister	4,82E-243
	Mitsuokella	8,24E-178
	Methanobrevibacter	3,20E-109
	Eggerthella	2,80E-65
	Ruminococcaceae	2,92E-63
	Subdoligranulum	3,40E-48
	Coprococcus	7,03E-31
	Collinsella	5,66E-29
	Blautia	1,92E-18
	Eubacterium	5,01E-15
	Dorea	0,000155341

No âmbito funcional foram identificados vários grupos KO e Módulos super-representados nos três enterotipos. Dentre eles observamos alguns relacionados à produção de vitaminas. Embora todas as vias de metabolismo de vitaminas estejam presentes em todas as amostras, os enterotipos 1 e 2 estão super-representados pela biossíntese de diferentes vitaminas. No grupo 1 observamos a prevalência de biotina, como foi visto na Figura 33 (B), riboflavina, pantotenato and ascorbato. Enquanto no grupo 2 temos a super-representação da biossíntese de tiamina, como mostrado na Figura 33 (C), e folato. A Tabela 17 mostra essa superrepresentação das vitaminas e outros módulos de vias, e seus respectivos valores-p.

Tabela 17 - Módulos de vias super-representados nos enterotipos e seus valores-p.

Enterotipo	Módulo	Descrição	Valor-p
1	M00155	Keratan sulfate degradation	4.48E-24
	M00006	Pentose phosphate pathway, oxidative phase	2.08E-15
	M00306	Sulfur reduction, sulfate => H2S	8.61E-12
	M00248	**Biotin biosynthesis, pimeloyl-CoA => biotin**	**2.85E-10**
	M00159	Fatty acid biosynthesis, initiation	1.58E-08
	M00008	Entner-Doudoroff pathway	1.87E-06
	M00001	Glycolysis, fructose-6P => pyruvate	2.53E-06
	M00017	Glutamate biosynthesis, oxoglutarate => glutamate (glutamate dehydrogenase)	3.90E-05
	M00244	**Pantothenate biosynthesis, valine => pantothenate**	**0.0016**
	M00250	**Riboflavin biosynthesis, GTP => riboflavin/FMN/FAD**	**0.0016**
	M00003	Gluconeogenesis, oxaloacetate => fructose-6P	0.0025
	M00608	PTS system, beta-glucosides-specific II component	0.0083
	M00255	**Ascorbate biosynthesis, animals**	**0.0148**
2	M00105	dTDP-Glucose, dTDP-galactose and dTDP-rhamnose biosynthesis	1.11E-29
	M00099	GDP-Mannose biosynthesis, fructose-6P => GDP-Man	2.32E-18
	M00252	**Thiamine biosynthesis, AIR => thiamine-P/thiamine-2P**	**5.37E-18**
	M00239	Ascorbate biosynthesis, plants	1.95E-16
	M00034	Tryptophan biosynthesis, chorismate => tryptophan	2.59E-16
	M00032	Cysteine biosynthesis, serine => cysteine	1.67E-06
	M00278	Complex II (succinate dehydrogenase / fumarate reductase), succinate dehydrogenase	4.68E-05
	M00042	Urea cycle	9.52E-05
	M00062	Cysteine metabolism, cysteine => 3-sulfino-L-alanine => pyruvate	0.0005
	M00299	C4-dicarboxylic acid cycle, phosphoenolpyruvate carboxykinase type	0.0007
	M00192	C5 isoprenoid biosynthesis, non-mevalonate pathway	0.0007
	M00063	Cysteine metabolism, cysteine => 3-mercaptopyruvate => pyruvate	0.0011
	M00240	NAD biosynthesis, aspartate => NAD	0.0047
	M00018	Glutamine biosynthesis, glutamate => glutamine	0.015
	M00022	Asparagine biosynthesis, aspartate => asparagine	0.0206
	M00029	Isoleucine biosynthesis, pyruvate => isoleucine	0.0283
	M00300	C4-dicarboxylic acid cycle, NAD+ -malic enzyme type	0.0316
	M00269	**Tetrahydrofolate biosynthesis**	**0.0471**
3	M00369	Peptides/nickel transport system	7.70E-111
	M00351	Simple sugar transport system	1.29E-62
	M00375	Cobalt transport system	8.47E-57
	M00376	Nickel transport system	3.49E-56
	M00367	Branched-chain amino acid transport system	1.99E-50
	M00337	Multiple sugar transport system	6.71E-47
	M00342	Ribose transport system	6.06E-20
	M00318	Sulfonate/nitrate/taurine transport system	2.20E-17
	M00353	Phosphonate transport system	1.15E-13
	M00320	Iron(III) transport system	1.49E-12
	M00309	Ribosome, archaea	7.94E-12
	M00303	Methane oxidation, methylotroph, methane => CO2	1.01E-09
	M00308	Ribosome, bacteria	3.51E-08
	M00246	**Heme biosynthesis, glutamate => protoheme/siroheme**	**2.11E-06**

5 DISCUSSÃO

As bases de dados biológicos são fontes de informações muito utilizadas para as pesquisas científicas, desempenhando papel fundamental na propagação do conhecimento a novas sequências e a análises sistêmicas. Entretanto, por muitas vezes, bases de dados de ortólogos sofrem com a falta de informação causada pela metodologia de criação ou pela falta de atualização, que não consegue acompanhar a gigantesca geração de novos dados.

Observamos que a base de dados KEGG Orthology, apesar da quantidade de dados já presentes, ainda tem muitas informações a serem acrescentadas. Isso foi evidenciado ao identificarmos uma grande proporção de anotações especuladas que foram solucionadas ao usarmos proteínas fora do KO para designar essas ESTs. Ou essas proteínas às quais as ESTs são designadas pertencem a agrupamentos do KO que não recrutaram proteínas de *C. elegans*, ou são proteínas relacionadas às que anotam a EST (o melhor alinhamento com a proteína do organismo que anotou também estaria na porção extra-KO) - neste último caso, a anotação já se caracteriza como um passo inicial para o estudo funcional completo de um pos´sivel novo grupo KO. Porém, o processo de agrupamento demanda uma intensa análise manual. Hoje, atendendo a necessidade de automatização do processo, a base KEGG já desenvolve um método, chamado KOALA, para propagar a anotação e aumentar os grupos de ortólogos já existentes (Kanehisa, Goto *et al.*, 2010).

O nosso procedimento de enriquecimento baseado no recrutamento de proteínas do UniProt, é um importante esforço para essa nova tendência de automações da propagação do conhecimento. Uma base de dados enriquecida confere ao usuário uma maior amplitude de informações, seja em níveis taxonômicos ou protéico, e para algumas áreas da ciência essa quantidade de dados torna-se essencial.

Um papel importante desempenhado pelo procedimento do UECOG é a criação de uma base chamada COG Editado. O COG não é atualizado desde 2003, ao longo desse período várias entradas deixaram de existir e vários GIs associados a proteínas COG foram substituídos por outros. A base COG Editado atualiza essas informações de GIs que foram trocados e remove entradas descontinuadas, além da remoção de fragmentos de proteínas presentes na base original. O procedimento de enriquecimento, por sua vez, mostra-se eficiente ao objetivo proposto, que é de aumentar a quantidade e qualidade da informação disponível nessa base de dados.

O enriquecimento da base UEKO mostrou uma enorme contribuição para o aumento

da quantidade de informações no KO. O aumento da quantidade de organismos representados é também de suma importância já que permite uma anotação de uma sequência por uma proteína de um organismo mais próximo. Essa disparidade filogenética entre KO e UEKO foi evidenciada ao compararmos os organismos usados pelas bases de dados para anotar ESTs e metagenomas. Nessa comparação, muitas vezes uma sequência era anotada por organismos distantes um do outro, compartilhando um ancestral comum no nível de super-reino. Ou seja, ao contrário até do que esperávamos, a melhoria da anotação não foi evidenciada somente pela adição de organismos mais próximos filogeneticamente aos já existentes na base. Outra melhoria qualitativa está mostrada na melhoria significativa dos escores das associações feitas pelo BLAST. Uma pequena diferença quantitativa também ocorreu. Registramos mais de 100 grupos KO com alinhamentos com sequencias de metagenômica, o que se refletiu em preenchimentos de algumas lacunas no mapa metabólico.

As bases UECOG e UEKO, apesar de utilizarem o mesmo procedimento para o enriquecimento, apresentam sequências e até mesmo grupos exclusivos. Cerca de 17,7% das entradas presentes no UECOG não estão presentes no UEKO. Dentre os 4.588 grupos presentes na versão UECOG 2.0, 1.160 não tem nenhuma de suas proteínas representadas no UEKO. Isso acontece porque 1.004 desses grupos que não estão incluídos no UEKO são formados por proteínas com função desconhecida, proteínas não caracterizadas ou com função geral predita. Entretanto existem informações relevantes que não estão presentes no UEKO, por exemplo, o grupo COG3209 (*Rhs family protein*), que está presente no UECOG com 473 entradas, e nenhuma delas está na base KO. Genes *RhsB*, que compõe o grupo no UECOG são encontrados na base KEGG Genes, mas não organizados em grupos de ortólogos.

Essa falta de informação na base de dados recrutadora passa a ser um fator limitante para o recrutamento final. Como exemplo temos um grupo K01123 (*sphingomyelin phosphodiesterase*), que não possui nenhuma proteína. Entretanto, o número EC associado a esse KO (EC: 3.1.4.41), é atribuído à 214 proteínas do UniProtKB, que teriam potencial para enriquecer a base caso esse grupo tivesse alguma entrada recrutadora. Caso semelhante acontece com os números EC: 2.1.1.57 (109 proteínas no UniProt) e EC: 2.7.7.68 (89 entradas UniProt) que não possuem um agrupamento KO para alocar suas proteínas.

Além da falta de proteínas recrutadoras, sofremos com a qualidade das mesmas quando erroneamente presentes em um grupo. Como exemplos temos as entradas do UniProtKB A1W7V7 e A1W9A1. Essas proteínas estão presentes na base KO originalmente no agrupamento K00134 (gliceraldeído 3-fosfato desidrogenase), mas na realidade pertencem

ao K00150 (gliceraldeído 3-fosfato desidrogenase NAD(P)). O procedimento UEKO é capaz de associar essas duas entradas ao K00150, porém não as retira do K00134. Um processo de edição e validação da base de dados original melhoraria a qualidade e quantidade do enriquecimento.

A análise metagenômica utilizando 13 amostras públicas permitiu uma valorização da qualidade da anotação proporcionada pela base UEKO. Além disso identificamos um genoma mínimo, composto por funções que todos os indivíduos tem, supostamente essencial para a sobrevivência da micriobiota. Nesse genoma mínimo identificamos funções exclusivas de infantes, e a evidente diferença da composição filogenética e funcional desses para com os indivíduos com microbiota já estável. Essa diferença orientada pela presença de bactérias do gênero *Bifidobacterium* já é conhecida e serve como suporte para a eficácia da nossa análise (Sela, Chapman *et al.*, 2008).

Uma outra análise metagenômica realizada identificou perfis filogenéticos interligados com perfis funcionais na microbiota. Esses perfis são chamados de enterotipos, e 2 deles são proeminentes e presentes em diferentes conjuntos de amostras estudadas. Esses perfis independem da etnia e hábitos alimentares dos indivíduos, uma vez que os enterotipos misturam amostras de diferentes países: Japoneses, Francês e Espanhol no grupo predominado por *Bacteroides*; e Espanhóis, Dinamarqueses, Francês e Italiano no grupo guiado por *Prevotella*. Além disso os enterotipos foram mantidos quando comparamos indivíduos do mesmo país.

É relevante ressaltar a similaridade entre os perfis filogenéticos e funcionais, e essa consistência dos enterotipos sugere que eles são resultado de uma composição definida e bem equilibrada da comunidade microbiana. Esses grupos podem ser usados para a identificação de um indivíduo, uma vez que estudos mostram um significante estabilidade da microbiota e que pode até ser restaurada após uma perturbação (Tannock, Munro *et al.*, 2000; Vanhoutte, Huys *et al.*, 2004; Costello, Lauber *et al.*, 2009).

Uma análise mais detalhada permite ligar as composições filogenética e funcional. Os principais gêneros do primeiro enterotipo, *Bacteroides* e com co-ocorrência de *Parabacteroides*, parecem ter sua fonte de energia primariamente através da fermentação de carboidratos e proteínas, uma vez que esses gêneros tem um imenso potencial sacarolítico (Martens, Koropatkin *et al.*, 2009). Além disso, genes que codificam enzimas envolvidas na degradação desses substratos (galactosidases, hexosaminidases, proteases), assim como glicólise e via de pentose-fosfato estão super-representadas, como mostrado na Tabela 17. A prevalência de *Prevotella* e *Desulfovibrio* no segundo enterotipo pode estar associada a

degradação de mucina, uma glicoproteína presente na mucosa intestinal. Esses dois gêneros agem em sinergia, uma vez que *Prevotella* é conhecido como um degradador de mucina e *Desulfovibrio* poderia melhorar o desempenho da degradação pela remoção do grupo sulfato (Wright, Rosendale *et al.*, 2000). O terceiro grupo é rico nos gêneros *Ruminococcus* e *Akkermansia*, também conhecidos pela degradação de mucina (Derrien, Vaughan *et al.*, 2004), e super-representado por transportadores de membrana, principalmente açúcares, sugerindo uma degradação da mucina e absorção dos açúcares simples resultantes.

O enriquecimento de determinados gêneros sugere que os enterotipos utilizam diferentes caminhos para a obtenção de energia fermentando substratos disponíveis no intestino, mostrando uma especialização e adaptação a diferentes nichos ecológicos.

A microbiota também tem um papel fundamental na conversão de substratos complexos em produtos que possam ser absorvidos por humanos, inclusive na produção de vitaminas. Observamos uma evidente complementaridade entre a microbiota e o metabolismo humano, dando mais suporte à esse visível mutualismo existente.

A aplicação das bases de dados enriquecidas em estudos funcionais reforça a necessidade de sua atualização de forma automatizada, tanto para estudos de genômica funcional, pois pesquisadores com interesses pontuais obtém uma amostragem de todas as proteínas completas com possível relação filogenética, o que foi exemplificado pela ocorrência de entradas de clados distantes, como também para análises sistêmicas, por sua vez exemplificado pela aplicação na caracterização metabólica. Os produtos UECOG e UEKO serão a partir daqui distribuídos pelo laboratório de Biodados da UFMG, em uma periodicidade mensal. Sua utilização, entretanto, em outros estudos do grupo de pesquisa já é frequente.

6 CONCLUSÕES

As análises evidenciaram algumas limitações das bases de dados estudadas por não conseguirem lidar com o grande volume de informação sendo gerada. Com isso temos a ausência de informações relevantes sobre vias e até mesmo sobre alguns organismos. A incapacidade de lidar com todos os dados gerados leva também a erros de classificação como pode ser observados em alguns exemplos.

Diante desse cenário, uma ferramenta de propagação do conhecimento já existente em bases de dados faz-se necessária. A integração de bases de dados e o desenvolvimento das ferramentas e metodologias para o enriquecimento da informação contida nessas bases é uma solução plausível para essa limitação.

No caso de bases desatualizadas como o COG, necessitamos de uma edição e atualização da informação contida. Mas essa edição é também necessária para qualquer base de dados que venha a ser utilizada como referência para um enriquecimento, umas vez que essa etapa eliminaria e filtraria informações incorretas ou desatualizadas.

As bases criadas, UEKO e UECOG, utilizam como referências diferentes fontes de informação (KEGG e COG, respectivamente), e por isso são complementares e o uso de uma não descarta o uso da outra. Isso acontece porque existem algumas funções que são caracterizadas exclusivamente em apenas uma das bases, caracterizando assim mais uma evidência da falta de informações. Muitos grupos de ortólogos ainda podem ser criados com informações de bases como o UniProt. Existem muitas entradas provenientes do Swiss-Prot com número EC caracterizado que poderia ser facilmente alocadas em um agrupamento KO, mas a ausência de um grupo para essa determinada função impede o enriquecimento da base original.

Essa maior quantidade de entradas protéicas e de organismos é de suma importância quando estudamos amostras ambientais, uma vez que a porção desconhecida de funções e organismos ali presentes é maior do que a porção já conhecida.

As bases UEKO e UECOG são utilizadas hoje para estudos evolutivos e análises contendo vias metabólicas de organismos que ainda não possuam o seu genoma completo.

7 REFERÊNCIAS BIBLIOGRÁFICAS

ADAMS, M. D. et al. Complementary DNA sequencing: expressed sequence tags and human genome project. **Science,** v. 252, n. 5013, p. 1651-6, Jun 21 1991.

ALTSCHUL, S. F. et al. Basic local alignment search tool. **J Mol Biol,** v. 215, n. 3, p. 403-10, Oct 5 1990.

ALTSCHUL, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Res,** v. 25, n. 17, p. 3389-402, Sep 1 1997.

AMANN, R. I.; LUDWIG, W.; SCHLEIFER, K. H. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. **Microbiol Rev,** v. 59, n. 1, p. 143-69, Mar 1995.

APWEILER, R. et al. UniProt: the Universal Protein knowledgebase. **Nucleic Acids Res,** v. 32, n. Database issue, p. D115-9, Jan 1 2004.

ARUMUGAM, M. et al. SmashCommunity: a metagenomic annotation and analysis tool. **Bioinformatics,** v. 26, n. 23, p. 2977-8, Dec 1 2010.

ARUMUGAM, M. et al. Enterotypes of the human gut microbiome. **Nature,** v. In Press, 2011.

BALL, C. A. et al. Standards for microarray data. **Science,** v. 298, n. 5593, p. 539, Oct 18 2002.

BARBOSA-SILVA, A. et al. **Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence**. BMC Bioinformatics. 9**:** 141 p. 2008.

BARRELL, D. et al. The GOA database in 2009--an integrated Gene Ontology Annotation resource. **Nucleic Acids Res,** v. 37, n. Database issue, p. D396-403, Jan 2009.

BATEMAN, A. et al. The Pfam protein families database. **Nucleic Acids Res,** v. 30, n. 1, p. 276-80, Jan 1 2002.

BATZOGLOU, S. et al. ARACHNE: a whole-genome shotgun assembler. **Genome Res,** v. 12, n. 1, p. 177-89, Jan 2002.

BOGUSKI, M. S.; TOLSTOSHEV, C. M.; BASSETT, D. E., JR. Gene discovery in dbEST. **Science,** v. 265, n. 5181, p. 1993-4, Sep 30 1994.

BORK, P.; KOONIN, E. V. Predicting functions from protein sequences--where are the bottlenecks? **Nat Genet,** v. 18, n. 4, p. 313-8, Apr 1998.

BOUCK, A.; VISION, T. The molecular ecologist's guide to expressed sequence tags.

Mol Ecol, v. 16, n. 5, p. 907-24, Mar 2007.

BRYANT, D. A.; FRIGAARD, N. U. Prokaryotic photosynthesis and phototrophy illuminated. **Trends Microbiol,** v. 14, n. 11, p. 488-96, Nov 2006.

CALINSKI, R. B.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics,** v. 3, p. 1-27, 1974.

CASPI, R. et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. **Nucleic Acids Res,** v. 36, n. Database issue, p. D623-31, Jan 2008.

CHEN, F. et al. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. **Nucleic Acids Res,** v. 34, n. Database issue, p. D363-8, Jan 1 2006.

CONSORTIUM, U. The Universal Protein Resource (UniProt) 2009. **Nucleic Acids Res,** v. 37, n. Database issue, p. D169-74, Jan 2009.

COSTELLO, E. K. et al. Bacterial community variation in human body habitats across space and time. **Science,** v. 326, n. 5960, p. 1694-7, Dec 18 2009.

DERRIEN, M. et al. Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. **Int J Syst Evol Microbiol,** v. 54, n. Pt 5, p. 1469-76, Sep 2004.

FERNANDES, G. R. et al. **A procedure to recruit members to enlarge protein family databases--the building of UECOG (UniRef-Enriched COG Database) as a model**. Genet Mol Res. 7**:** 910-24 p. 2008.

FERNANDES, G. R.; MUDADO, M. A.; ORTEGA, J. M. **Testing the performance of automated annotation of ESTs with the Kegg Orthology (KO) database demonstrates lack of completeness of clusters**. Genet Mol Res. 7**:** 948-57 p. 2008.

FITCH, W. M. Distinguishing homologous from analogous proteins. **Syst Zool,** v. 19, n. 2, p. 99-113, Jun 1970.

FLEISCHMANN, R. D. et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. **Science,** v. 269, n. 5223, p. 496-512, Jul 28 1995.

GILL, S. R. et al. Metagenomic analysis of the human distal gut microbiome. **Science,** v. 312, n. 5778, p. 1355-9, Jun 2 2006.

HANDELSMAN, J. et al. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. **Chem Biol,** v. 5, n. 10, p. R245-9, Oct 1998.

HUANG, X.; MADAN, A. CAP3: A DNA sequence assembly program. **Genome Res,** v. 9, n. 9, p. 868-77, Sep 1999.

HUANG, Y.; GILNA, P.; LI, W. Identification of ribosomal RNA genes in metagenomic fragments. **Bioinformatics,** v. 25, n. 10, p. 1338-40, May 15 2009.

HUGENHOLTZ, P.; GOEBEL, B. M.; PACE, N. R. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. **J Bacteriol,** v. 180, n. 18, p. 4765-74, Sep 1998.

HUNTER, S. et al. InterPro: the integrative protein signature database. **Nucleic Acids Res,** v. 37, n. Database issue, p. D211-5, Jan 2009.

HUSON, D. H. et al. **MEGAN analysis of metagenomic data**. Genome Res. 17**:** 377-86 p. 2007.

JENSEN, L. J. et al. eggNOG: automated construction and annotation of orthologous groups of genes. **Nucleic Acids Res,** v. 36, n. Database issue, p. D250-4, Jan 2008.

KANEHISA, M. A database for post-genome analysis. **Trends Genet,** v. 13, n. 9, p. 375-6, Sep 1997.

KANEHISA, M.; BORK, P. Bioinformatics in the post-sequence era. **Nat Genet,** v. 33 Suppl, p. 305-10, Mar 2003.

KANEHISA, M. et al. KEGG for representation and analysis of molecular networks involving diseases and drugs. **Nucleic Acids Res,** v. 38, n. Database issue, p. D355-60, Jan 2010.

KANEHISA, M. et al. The KEGG resource for deciphering the genome. **Nucleic Acids Res,** v. 32, n. Database issue, p. D277-80, Jan 1 2004.

KAUFMAN, L.; J. ROUSSEEUW, P. **Finding groups in data: an introduction to cluster analysis:** 342 p. 2005.

KITANO, H. et al. Using process diagrams for the graphical representation of biological networks. **Nat Biotechnol,** v. 23, n. 8, p. 961-6, Aug 2005.

KROGFELT, K. A. Bacterial adhesion: genetics, biogenesis, and role in pathogenesis of fimbrial adhesins of Escherichia coli. **Rev Infect Dis,** v. 13, n. 4, p. 721-35, Jul-Aug 1991.

KROGH, A. et al. Hidden Markov models in computational biology. Applications to protein modeling. **J Mol Biol,** v. 235, n. 5, p. 1501-31, Feb 4 1994.

KUROKAWA, K. et al. **Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes**. DNA Res. 14**:** 169-81 p. 2007.

LANDER, E. S. et al. Initial sequencing and analysis of the human genome. **Nature,** v. 409, n. 6822, p. 860-921, Feb 15 2001.

LARKIN, M. A. et al. Clustal W and Clustal X version 2.0. **Bioinformatics,** v. 23, n. 21, p. 2947-8, Nov 1 2007.

LETUNIC, I.; BORK, P. **Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation**. Bioinformatics. 23**:** 127-8 p. 2007.

LETUNIC, I. et al. iPath: interactive exploration of biochemical pathways and networks. **Trends Biochem Sci,** v. 33, n. 3, p. 101-3, Mar 2008.

LI, W.; GODZIK, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. **Bioinformatics,** v. 22, n. 13, p. 1658-9, Jul 1 2006.

LOYTYNOJA, A.; GOLDMAN, N. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. **BMC Bioinformatics,** v. 11, p. 579, 2010.

LUKASHIN, A. V.; BORODOVSKY, M. GeneMark.hmm: new solutions for gene finding. **Nucleic Acids Res,** v. 26, n. 4, p. 1107-15, Feb 15 1998.

MAO, X. et al. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. **Bioinformatics,** v. 21, n. 19, p. 3787-93, Oct 1 2005.

MARCHLER-BAUER, A. et al. CDD: a Conserved Domain Database for protein classification. **Nucleic Acids Res,** v. 33, n. Database issue, p. D192-6, Jan 1 2005.

MARGULIES, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. **Nature,** v. 437, n. 7057, p. 376-80, Sep 15 2005.

MARTENS, E. C. et al. Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm. **J Biol Chem,** v. 284, n. 37, p. 24673-7, Sep 11 2009.

MATSUOKA, Y. et al. Payao: a community platform for SBML pathway model curation. **Bioinformatics,** v. 26, n. 10, p. 1381-3, May 15 2010.

MEYER, F. et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. **BMC Bioinformatics,** v. 9, p. 386, 2008.

MILLIGAN, G.; COOPER, M. An examination of procedures for determining the number of clusters in a data set. **Psychometrika,** v. 2, p. 159-179, 1985.

MITRA, R. D.; CHURCH, G. M. In situ localized amplification and contact replication of many individual DNA molecules. **Nucleic Acids Res,** v. 27, n. 24, p. e34, Dec 15 1999.

MITRA, R. D. et al. Fluorescent in situ sequencing on polymerase colonies. **Anal Biochem,** v. 320, n. 1, p. 55-65, Sep 1 2003.

MORIYA, Y. et al. KAAS: an automatic genome annotation and pathway reconstruction server. **Nucleic Acids Res,** v. 35, n. Web Server issue, p. W182-5, Jul 2007.

MORTAZAVI, A. et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. **Nat Methods,** v. 5, n. 7, p. 621-8, Jul 2008.

MUDADO, M. A.; ORTEGA, J. M. A picture of gene sampling/expression in model organisms using ESTs and KOG proteins. **Genet Mol Res,** v. 5, n. 1, p. 242-53, 2006.

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **J Mol Biol,** v. 48, n. 3, p. 443-53, Mar 1970.

NELSON, K. E. et al. A catalog of reference genomes from the human microbiome. **Science,** v. 328, n. 5981, p. 994-9, May 21 2010.

PERTEA, G. et al. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. **Bioinformatics,** v. 19, n. 5, p. 651-2, Mar 22 2003.

QIN, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. **Nature,** v. 464, n. 7285, p. 59-65, Mar 4 2010.

RAES, J.; BORK, P. Molecular eco-systems biology: towards an understanding of community function. **Nat Rev Microbiol,** v. 6, n. 9, p. 693-9, Sep 2008.

RAES, J.; FOERSTNER, K. U.; BORK, P. Get the most out of your metagenome: computational analysis of environmental sequence data. **Curr Opin Microbiol,** v. 10, n. 5, p. 490-8, Oct 2007.

RAPPE, M. S.; GIOVANNONI, S. J. The uncultured microbial majority. **Annu Rev Microbiol,** v. 57, p. 369-94, 2003.

SALONEN, A. et al. Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. **J Microbiol Methods,** v. 81, n. 2, p. 127-34, May 2010.

SANGER, F. Determination of nucleotide sequences in DNA. **Biosci Rep,** v. 1, n. 1, p. 3-18, Jan 1981.

SANGER, F. et al. Nucleotide sequence of bacteriophage phi X174 DNA. **Nature,** v. 265, n. 5596, p. 687-95, Feb 24 1977.

SAYERS, E. W. et al. Database resources of the National Center for Biotechnology Information. **Nucleic Acids Res,** v. 39, n. Database issue, p. D38-51, Jan 2011.

SELA, D. A. et al. The genome sequence of Bifidobacterium longum subsp. infantis reveals adaptations for milk utilization within the infant microbiome. **Proc Natl Acad**

Sci U S A, v. 105, n. 48, p. 18964-9, Dec 2 2008.

SESHADRI, R. et al. CAMERA: a community resource for metagenomics. **PLoS Biol,** v. 5, n. 3, p. e75, Mar 2007.

SHENDURE, J.; JI, H. Next-generation DNA sequencing. **Nat Biotechnol,** v. 26, n. 10, p. 1135-45, Oct 2008.

SHENDURE, J. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. **Science,** v. 309, n. 5741, p. 1728-32, Sep 9 2005.

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. **J Mol Biol,** v. 147, n. 1, p. 195-7, Mar 25 1981.

STEIN, L. D. Integrating biological databases. **Nat Rev Genet,** v. 4, n. 5, p. 337-45, May 2003.

SUZEK, B. E. et al. UniRef: comprehensive and non-redundant UniProt reference clusters. **Bioinformatics,** v. 23, n. 10, p. 1282-8, May 15 2007.

TANNOCK, G. W. et al. Analysis of the fecal microflora of human subjects consuming a probiotic product containing Lactobacillus rhamnosus DR20. **Appl Environ Microbiol,** v. 66, n. 6, p. 2578-88, Jun 2000.

TATUSOV, R. L. et al. The COG database: an updated version includes eukaryotes. **BMC Bioinformatics,** v. 4, p. 41, Sep 11 2003.

TATUSOV, R. L.; KOONIN, E. V.; LIPMAN, D. J. A genomic perspective on protein families. **Science,** v. 278, n. 5338, p. 631-7, Oct 24 1997.

THOMAS, P. D.; MI, H.; LEWIS, S. Ontology annotation: mapping genomic regions to biological function. **Curr Opin Chem Biol,** v. 11, n. 1, p. 4-11, Feb 2007.

TRINGE, S. G. et al. Comparative metagenomics of microbial communities. **Science,** v. 308, n. 5721, p. 554-7, Apr 22 2005.

TURNBAUGH, P. J. et al. A core gut microbiome in obese and lean twins. **Nature,** v. 457, n. 7228, p. 480-4, Jan 22 2009.

VANHOUTTE, T. et al. Temporal stability analysis of the microbiota in human feces by denaturing gradient gel electrophoresis using universal and group-specific 16S rRNA gene primers. **FEMS Microbiol Ecol,** v. 48, n. 3, p. 437-46, Jun 1 2004.

VELCULESCU, V. E. et al. Serial analysis of gene expression. **Science,** v. 270, n. 5235, p. 484-7, Oct 20 1995.

VENTER, J. C. et al. The sequence of the human genome. **Science,** v. 291, n. 5507, p. 1304-51, Feb 16 2001.

VENTER, J. C. et al. Environmental genome shotgun sequencing of the Sargasso

Sea. **Science,** v. 304, n. 5667, p. 66-74, Apr 2 2004.

WANG, Q. et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. **Appl Environ Microbiol,** v. 73, n. 16, p. 5261-7, Aug 2007.

WREN, J. D.; BATEMAN, A. Databases, data tombs and dust in the wind. **Bioinformatics,** v. 24, n. 19, p. 2127-8, Oct 1 2008.

WRIGHT, D. P.; ROSENDALE, D. I.; ROBERTSON, A. M. Prevotella enzymes involved in mucin oligosaccharide degradation and evidence for a small operon of genes expressed during growth on mucin. **FEMS Microbiol Lett,** v. 190, n. 1, p. 73-9, Sep 1 2000.

8 APÊNDICE

Os trabalhos desenvolvidos para esta tese geraram publicações e referências para trabalhos do nosso grupo e de outros grupos de pesquisa.

FERNANDES, G. R. et al. **A procedure to recruit members to enlarge protein family databases--the building of UECOG (UniRef-Enriched COG Database) as a model**. Genet Mol Res. 7**:** 910-24 p. 2008.

FERNANDES, G. R.; MUDADO, M. A.; ORTEGA, J. M. **Testing the performance of automated annotation of ESTs with the Kegg Orthology (KO) database demonstrates lack of completeness of clusters**. Genet Mol Res. 7**:** 948-57 p. 2008.

ARUMUGAM, M. et al. Enterotypes of the human gut microbiome. **Nature,** v. In Press, 2011.

Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

A procedure to recruit members to enlarge protein family databases - the building of UECOG (UniRef-Enriched COG Database) as a model

G.R. Fernandes[1]*, D.V.C. Barbosa[1]*, F. Prosdocimi[1], I.A. Pena[1], L. Santana-Santos[1], O. Coelho Junior[1], A. Barbosa-Silva[1], H.M. Velloso[1], M.A. Mudado[2], D.A. Natale[3], A.C. Faria-Campos[4], S.V. A. Campos[4] and J.M. Ortega[1]

[1]Departamento de Bioquímica e Imunologia, Laboratório de Biodados, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
[2]Fundação Ezequiel Dias, Belo Horizonte, MG, Brasil
[3]Protein Information Resource, Georgetown University Medical Center, Washington, DC, USA
[4]Departamento de Ciência da Computação, ICEX, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
*These authors contributed equally to this study.

Corresponding author: J.M. Ortega
E-mail: miguel@icb.ufmg.br

Genet. Mol. Res. 7 (3): 910-924 (2008)
Received June 2, 2008
Accepted August 11, 2008
Published September 30, 2008

ABSTRACT. A procedure to recruit members to enlarge protein family databases is described here. The procedure makes use of UniRef50 clusters produced by UniProt. Current family entries are used to recruit additional members based on the UniRef50 clusters to which they belong. Only those additional UniRef50 members that are not fragments and whose length is within a restricted range relative to the original entry are recruited. The enriched dataset is then limited to contain only genomes from selected clades. We used the COG database - used for genome annotation and for studies of phylogenetics and gene evolution - as a model. To validate the method, a

UniRef-Enriched COG0151 (UECOG) was tested with distinct procedures to compare recruited members with the recruiters: PSI-BLAST, secondary structure overlap (SOV), Seed Linkage, COGnitor, shared domain content, and neighbor-joining single-linkage, and observed that the former four agree in their validations. Presently, the UniRef50-based recruitment procedure enriches the COG database for Archaea, Bacteria and its subgroups Actinobacteria, Firmicutes, Proteobacteria, and other bacteria by 2.2-, 8.0-, 7.0-, 8.8-, 8.7-, and 4.2-fold, respectively, in terms of sequences, and also considerably increased the number of species.

Key words: COG; Secondary database; UniRef; UniProt; UECOG

INTRODUCTION

Phylogenetics, genome annotation and studies of gene evolution all benefit from the comparative analysis of protein families. Several databases are dedicated to the clustering of related proteins. Indeed the COG database (Tatusov et al., 1997, 2003) - a collection of Clusters of Orthologous Groups (COGs) of proteins - has not only supported diverse analysis of the protein families, but has stimulated the development of databases derived from the use of distinct procedures, such as OrthoMCL-DB (Chen et al., 2006) and Inparanoid (O'Brien et al., 2005). In some sense, attribution of Gene Ontology terms to amino acid sequences by the GOA project (Camon et al., 2004) also forms collections of genes with the same activity in distinct organisms, therefore being putative orthologs. Other recent databases or procedures to cluster sequences also address the same issue, such as FlowerPower (Krishnamurthy et al., 2007), Ortholuge (Fulton et al., 2006), OrthologID (Chiu et al., 2006), and Seed Linkage (Barbosa-Silva et al., 2008). In addition to these, the Universal Protein Resource Consortium (UniProt) produces the UniProt Knowledgebase (UniProt Consortium, 2007) and the UniProt Reference Clusters (UniRef) containing over five million sequences. The latter set contains equivalents of protein families that are generated triweekly using CD-HIT (Li and Godzik, 2006), resulting in three distinct types of clusters: i) UniRef100, where the best representative of 100% identical entries is selected to stand for the sequences in that cluster; ii) UniRef90 and iii) UniRef50, where members of each cluster show either 90 or 50% identity, respectively, to the seed sequence (Suzek et al., 2007).

An attractive possibility would be to use UniRef50 clusters to enrich a database built with complete genomes such as COG. The rationale for doing so is that the UniRef50 clusters are generated triweekly, and thus provide a basis for rapid enrichment of less-frequently updated databases. Here, we describe the procedure to build UniRef-Enriched COGs (UECOGs) and present a case study using multiple validation procedures for recruitment. These latter procedures include: a) PSI-BLAST (position-specific iterated - basic local alignment search tool), where all recruited UniRef50 members that are hit under a PSI-BLAST (Altschul et al., 1997) search started by the recruiter (a COG member from the same UniRef50 cluster) are labeled as valid recruited entries; b) secondary structure overlap (SOV), where recruited members whose indices of secondary structure overlap as determined by SSPro4 (Geourjon et al., 2001) and SOV (Rost et al., 1994) are labeled as valid if over a given threshold; c) neighbor-joining tree neighboring, where all recruited entries that are in single linkage (continuously consecutive) to the recruiter are labeled as valid; d) domain structure, where recruited members that share the same content of domains as determined by RPS-BLAST using SMART (Schultz et al., 1998), Pfam (Finn et al., 2006) and COG domain databases,

and a threshold of 75% coverage and 1e-3 E-value cutoff are applied; e) COGnitor at the NCBI website, using the old COG version as database, determining if the recruited entries belong to the expected COG; f) Seed Linkage, a software developed by our group to enlarge clusters using their members as seed for recruitment of cognate proteins (Barbosa-Silva et al., 2008).

Presently, UECOG enriches the COG database for Archaea, Bacteria and its subgroups Actinobacteria, Firmicutes, Proteobacteria, and other bacteria by 2.2-, 8.0-, 7.0-, 8.8-, 8.7-, and 4.2-fold, respectively, in terms of sequence, and also increased the number of species. Users can download UECOG for the distinct clades from our server at http://biodados.icb.ufmg.br/uecog. UECOG is updated monthly using the latest available iProClass table and UniProtKB file.

MATERIAL AND METHODS

Enrichment of COG database with members from UniRef50 clusters

Two tables of tabulated data were used for this purpose: i) the COG file whog, downloaded from NCBI [ftp://ftp.ncbi.nlm.nih.gov/pub/COG/COG] and ii) iProClass (Huang et al., 2003) downloaded September 14, 2007 from the PIR FTP site [ftp://ftp.pir.georgetown.edu/databases/iproclass/]. The table obtained from COG was updated to contain the txid NCBI taxonomy information and the taxon group IDs or clade Archaea or Bacteria (which was further subdivided into Actinobacteria, Firmicutes, Proteobacteria, and other bacteria). From the COG table, the genbank identifier (*gi*) ID was extracted and the iProClass table was consulted to determine which COG members would act as recruiters. The UniRef50 cluster ID for recruiters was obtained. All members from the UniRef50 recruited clusters were then recruited.

Filtering of UniParc entries and fragments

Certain entries included in UniRef50 are not from the richly annotated set included in UniProtKB. Entries from the UniProt Archive (UniParc) - not useful for further analysis - were filtered out of the database based on the "UPI" prefix of the entry identifier. After the FASTA file for the other entries was obtained, a parser was used to inspect the annotation in FASTA file in order to filter out all those containing the string "(Fragment)", which denotes entries without a functional start or stop codon.

Filtering by taxonomic clade

One possibly undesirable effect of this enrichment procedure is the addition of eukaryotic sequences to the (largely prokaryotic) COG clusters. Thus, we decided to filter the recruitment using the taxon group (clade) ID. Clade subtrees can be obtained at the NCBI taxomomy site using, for example, the query "txid2 [Subtree]" and formatting the results as txid list. The final UECOG thus contains only Archaea and Bacteria subtrees. The eukaryotic organisms *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Encephalitozoon cuniculi* were not included in UECOG.

Filtering by size selection

The size of each potential recruit was compared to that of its recruiter. If the ratio

ranged from 0.9 to 1.1 - where the recruited protein was not more than 10% shorter or longer than the recruiter - the recruitment was allowed.

Validation by PSI-BLAST

The PSI-BLAST validation uses COG members as queries (recruiters) and cognate UniRef50 members (recruited) plus recruiters as a formatted database. The blastpgp program was performed with parameter -h 1 x 10^{-5} (an E-value cutoff that limits the inclusion of sequences in the alignment that is used to construct the position-specific scoring matrices, PSSM) and run to convergence. Recruited sequences hitting at least one recruiter were considered to be PSI-BLAST validated. The search was run either using the entire UECOG as database or concentrated on a specific clade (e.g., Proteobacteria).

Validation by domain conservation

Conserved domains were mapped to recruiters using RPS-BLAST with the CDD database (which contains SMART, Pfam and COG) (Marchler-Bauer et al., 2002). A stringent procedure was executed to map the domains, requiring an E-value lower than 1E-3 and over 75% coverage of the domain size in the CDD database. The occurrence of domains was examined in recruiters to determine the list of domains shared among all recruiters that are from the same UniRef50 cluster within a given COG. All recruited sequences from this same UniRef50 were considered domain validated if they showed the same domains shared by recruiters in the cognate UniRef50 cluster.

Validation by branch distance in neighbor-joining trees

Recruitment of UniRef50 members was evaluated by construction of neighbor-joining trees. Sequences from a given clade were aligned with Clustal W using BLOSUM62 and default parameters, submitted to SEQBOOT to generate five adjusted multiple alignments, and the distance between them was estimated with PHYLIP PROTDIST. Moreover, five trees were generated with PHYLIP NEIGHBOR, and the distance and number of branches between all proteins were calculated and stored. To determine the neighboring in a tree, a single-linkage procedure was started with the original COG members that acted as recruiters and the number of sequences clustered by an iterative one branch distance was determined. When the search reached a COG member, the count was not incremented but the search was continued. Sequences that were one branch apart in single-linkage iterations were considered neighbor-joining validated. Merging of UniRef50 clusters in neighbor-joining trees was the prominent cause of search interruption.

Validation by secondary structure overlap

The secondary structure of recruiters and recruited proteins was determined using a local implementation of the SSPro4 software. Percentage of structural overlap was determined with the SOV parameter described by Rost et al. (1994) for tuples of sequences from the same UniRef50 to determine the minimum SOV value amongst them. Recruiters were aligned with recruited candidates and the maximum SOV values were stored for each recruited sequence. In the case that a given UniRef50 member was the sole recruiter, SOV values of alignments of recruited sequences

to this recruiter were used. The results were processed by classes of SOV. Sequences that passed the 80% SOV against recruiters were considered SOV validated.

Validation by Seed Linkage

Recruiters from the Edited COG0151 were used as seed in the Seed Linkage program (Barbosa-Silva et al., 2008) using as a database the entire UECOG0151. No seed remained as a singlet and a single cluster was formed.

Availability

UECOG is available from our server at http://biodados.icb.ufmg.br/uecog. UECOG will be updated on a monthly basis. Future versions will incorporate web services.

RESULTS

The use of COGs for phylogenetic studies or as a source of information for the annotation of novel genomes would benefit from a reliable update. One example is given in Figure 1. Two strains of *Helicobacter pylori* are present in COGs, and often proteins from both of these strains are grouped as brothers in neighbor-joining trees (for example, see the branch labeled "A" in Figure 1, left panel), but sometimes an expansion is exclusive to a given genome (see branch "B", also in the left panel). The enrichment with additional sequences obtained from an updated dataset such as the UniRef50 database corroborates the observation that gene B is not shared equally between *H. pylori* genomes for strains 26695 and J99 (Figure 1, right panel). Thus, we set out to enrich COG with UniRef50 entries.

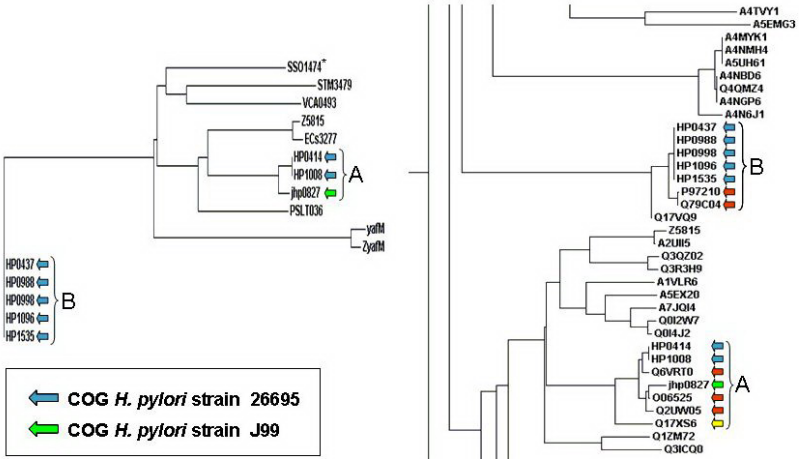

Figure 1. Neighbor-joining trees using COG or UECOG databases. On the left, a region of the phylogram showing two putatively distant transposases (COG1943), one (gene A) present in both *Helicobacter pylori* strains 26695 (blue arrows) and J99 (green arrows) and the other potentially exclusive to one of these strains (gene B). On the right, a region of the phylogram obtained with UECOG confirming this exclusivity after the addition of 154 Proteobacteria sequences that included sequences from other *H. pylori* strains (red arrows) and from *H. acinonychis* (yellow arrow). Sequence SSO1474* from *Sulfolobus solfataricus* (Archaea) was used to represent the root of the unrooted tree on the left.

Production of an updated version of COG

COG is the most used ortholog database for genome and gene annotation analyses. It was initially built in 1997 and further updated in 2001. UniProt produces a triweekly updated database containing well-annotated protein sequences from a plethora of organisms, and includes a clustered set called UniRef50. Our aim was the production of a UniRef-Enriched COG database to allow updated gene annotation based on ortholog groups and phylogenetic studies with a more complete source of information. In order to produce UECOG, we first produced an updated version of COG (Edited COG), since this database now contains some original sequences that are no longer valid. In this initial analysis, we identified 4506 NCBI *gi* in COG, which were discontinued. Of these invalid *gi*'s, 2091 could be mapped to new valid *gi*'s using Batch Entrez (http://www.ncbi.nlm.nih.gov/sites/batchentrez) that linked old *gi*'s to new ones based on accession number data. Moreover, to be maintained in Edited COG, all entries must have a corresponding accession in UniProtKB. To find the UniProtKB equivalent to those proteins that failed to return a UniProtKB accession by consulting iProClass with either its *gi* or the updated *gi*, we performed a BLAST search against the UniProtKB database using the amino acid sequence in COG as query. We analyzed the BLAST results to select proteins identical to the COG sequence; this UniProtKB entry was put into the Edited COG. After this procedure, we applied a last filter to delete the entries that were labeled as "Fragment" in the UniProtKB FASTA descriptions. The whole procedure is illustrated in Figure 2.

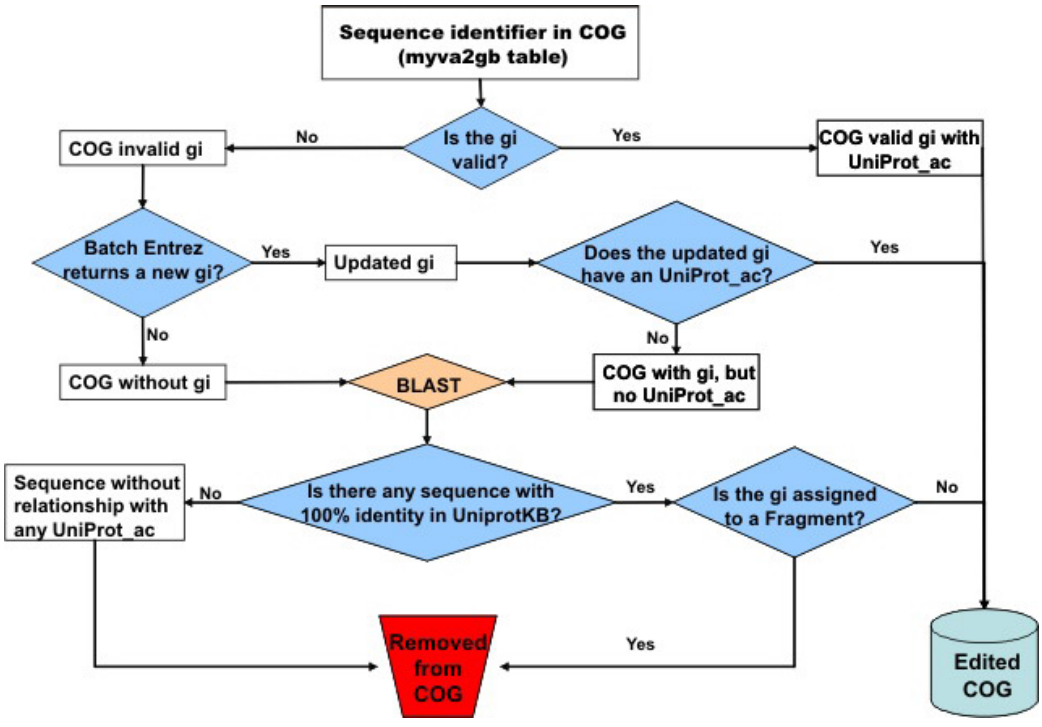

Figure 2. Methodology of COG edition. Discontinued *gi* in COG were updated with Batch Entrez or with BLAST to UniProt, and those entries without a reference in UniProt were removed from Edited COG.

Database recruitment procedure

All the Edited COG entries have a corresponding UniProtKB entry, and have thus been assigned to a UniRef50 cluster as well. The COG entries that are members of a UniRef50 cluster were called recruiters. In the next step, we selected all non-fragment UniProtKB entries that share a UniRef50 cluster with one recruiter protein. Each of these recruited proteins joined its recruiter COG cluster. All the steps are shown in Figure 3.

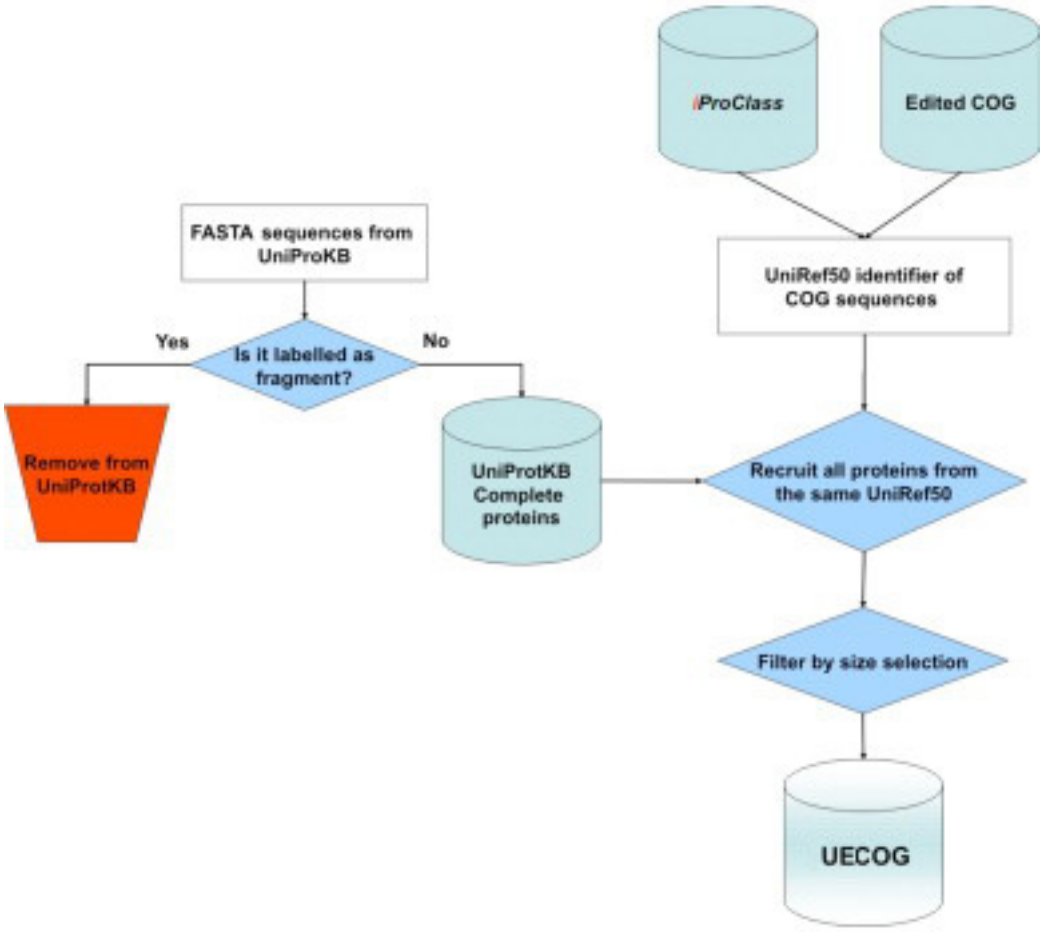

Figure 3. Methodology to build UECOG. Members of Edited COG acted as recruiters of members of their UniRef50 clusters, but only complete proteins were recruited. Size selection was further added at the recruitment step (see below).

Enrichment of COG database with members from UniRef50 clusters

UECOG enrichment was performed by recruiting non-fragment members of UniRef50 clusters that share similar length with one or more COG members (see Material and Methods).

Only prokaryotic organisms comprise the UECOG database. The three eukaryotic organisms were removed from Edited COG, because there is a specific database for the eukaryotic organisms (KOG), and the filtering by clade would not be easily determined. The final database, consisting of prokaryotic sequences, was enriched with 961,725 proteins (7.01-fold compared to the original COG database), as shown in Table 1. Sequences from Proteobacteria proved to be the greatest source of UECOG enrichment (8.74-fold), while Actinobacteria showed the greatest source of enrichment of species growing from 4 species in COG to 6736 in UECOG. Archaea yielded the smallest enrichment of sequence and genomes as compared to other groups, probably because it is very distantly related to other bacterium groups, and there is no great number of Archaea already sequenced.

Table 1. Enrichment in UECOG.

Clade	COG		Edited COG		UECOG		Fold
	Genomes	Proteins	Genomes	Proteins	Genomes	Proteins	
COG	66	144320	n/a	n/a	n/a	n/a	n/a
Prokaryotes	63	137122	63	124369	3477	961725	7.01
Archaea	13	22374	13	21310	248	49836	2.23
Bacteria	53	114748	50	103059	3229	911889	7.95
Actinobacteria	4	9391	4	6736	391	65871	7.01
Firmicutes	12	20921	12	19961	747	184403	8.81
Proteobacteria	24	67737	24	60741	1594	592000	8.74
Other bacteria	14	16699	10	15621	497	69615	4.17

n/a = not available.

Moreover, the overall enrichment was measured for each UECOG cluster. The majority of the COG clusters were enriched with new proteins. The data in Figure 4 present the amount of enrichment for each COG. Each dot represents a single COG. The original number of sequences for each COG was plotted on the X-axis while the enriched UECOG number is shown on the Y-axis. The dots above the dashed line represent the enriched clusters. The data show that some UECOG clusters are enriched up to about 30-fold.

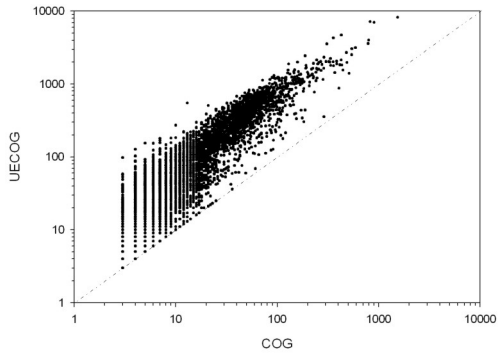

Figure 4. Enrichment in UECOG. The number of entries in UECOG is shown as a function of the entries in Edited COG. A dashed line indicates the cases where no enrichment was obtained.

Recruitment validation procedures

To validate the recruitment of UniRef50 proteins to each COG, we performed PSI-BLAST and RPS-BLAST searches, comparisons of SOV, and single linkage inspection of neighbor-joining trees, and used the Seed Linkage program (Barbosa-Silva et al., 2008) using recruiters as seed. The validation experiments described below were conducted on UECOGs constructed without size selection.

We chose for a case analysis the cluster COG0151 (phosphoribosylamine-glycine ligase) containing 53 proteins in COG; of these, the 51 prokaryote sequences (excluding *S. cerevisiae*, *S. pombe*, and *E. cuniculi*) were used as queries to illustrate the validation. The corresponding UECOG was selected because it was greatly enriched (with 501 proteins), where its protein sequences have broadly conserved domains and it possesses proteins from all clades of the 66 organisms of COG. Of the 501 candidates, 42 would be discarded by size filtering; however, they were retained in these experiments to illustrate the appropriateness of the filtering by size. With PSI-BLAST, the validation was executed in two flavors: validation of the whole UECOG as database or using six sub-sets (Archaea, Bacteria, Actinobacteria, Firmicutes, Proteobacteria, and other bacteria) as databases. PSI-BLAST is a BLAST module that, with iterative searches creates the PSSM related to query and the protein sequences that are found in each round. Therefore, PSI-BLAST searches can be more sensitive, finding more similar proteins and obtaining more orthologs than BLASTp. The searches were carried out using each recruiter protein (from COG) as query against its corresponding UECOG until the convergence of sequences was found. The recruited protein was considered "validated" when it was found as a hit by at least one recruiter.

In the search using the whole set, the 51 queries against the 552 proteins of UECOG0151 (recruited + recruiters) were used and 94% of recruited candidates were validated. Conversely, the six sub-sets were generated in accordance to the taxonomy clades and an analysis was performed to verify the differences in the performance of validation caused by the restriction of recruiters and recruited groups to a clade, thus resulting in a search focused on that clade. In this way, we could validate an equal or greater number of recruited candidates. The results are shown in Table 2 and indicate that validation by both means is rather high.

Table 2. PSI-BLAST, Seed Linkage, and secondary structure overlap (SOV) validation of UECOG0151.

Clade	Recruiter	Recruited	PSI-BLAST		Seed linkage	SOV
			UECOG database validated	Clade database validated	Validated	Validated
Prokaryotes	51	501	470 (94%)	470 (94%)	463 (92%)	470 (94%)
Archaea	12	30	30 (100%)	30 (100%)	29 (97%)	29 (97%)
Bacteria	41	471	438 (93%)	446 (95%)	434 (92%)	441 (94%)
Actinobacteria	4	41	38 (93%)	41 (100%)	40 (98%)	41 (100%)
Firmicutes	8	93	92 (99%)	92 (99%)	92 (99%)	92 (99%)
Proteobacteria	21	313	279 (89%)	286 (91%)	282 (90%)	288 (92%)
Other bacteria	6	24	20 (83%)	20 (83%)	20 (83%)	20 (83%)

Another validation procedure was conducted by the Seed Linkage software produced by our group (Barbosa-Silva et al., 2008). This software produces a clustering of cognate proteins from multiple organisms beginning with a single sequence through connectivity saturation with that seed sequence. Thus, recruiters were used as seed and a file containing both recruiters and recruited proteins was used as database. Seed Linkage returned a single cluster as expected, but excluded 38 sequences, distributed into diverse clades (Table 2).

One more validation procedure was applied by comparing secondary structure of recruited candidates to every recruiter from the same UniRef50 cluster, in a pair-wise fashion. First, secondary structures of all proteins were predicted by the SSPro4 software (Geourjon et al., 2001). Next, structural overlap between recruited candidates and recruiters was determined as in Rost et al. (1994), and the highest SOV index obtained was saved. Figure 5 shows the distribution of SOV in classes. Most of the recruited candidates show a SOV value over 80%, a value that splits the distribution into two groups. Thus, 80% SOV was used as cutoff for validation. The distribution of validation into distinct clades is shown in Table 2. Validation with SOV was also very high.

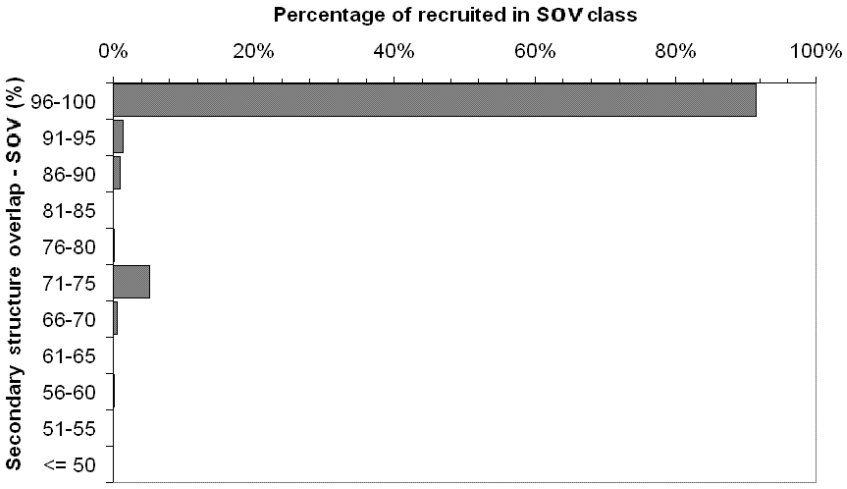

Figure 5. Distribution of secondary structure overlap (SOV) between recruited candidates and recruiters. SOV values were determined for each recruited candidate and the respective recruiters, and the highest value was considered.

An additional verification of the enrichment was conducted by posting the recruited candidates to a search in the COGnitor program available at the NCBI site (http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html). This analysis confirmed that 29 of the 30 entries that were not validated by three approaches (PSI-BLAST, SOV and Seed Linkage) actually belong to a different cluster, COG0041. The only sequence that COGnitor mapped to COG0151 actually is a fragment derived from genome annotation. A Venn diagram analyzing the four procedures is shown in Figure 6A. Twenty-nine sequences were not validated by the four procedures and most of the sequences (463 of 501: 92.4%) were four times validated. Figure 6B shows the diagram using only sequences that passed the size selection; with a cost of generating some false negatives, the data indicate that false positives were discarded.

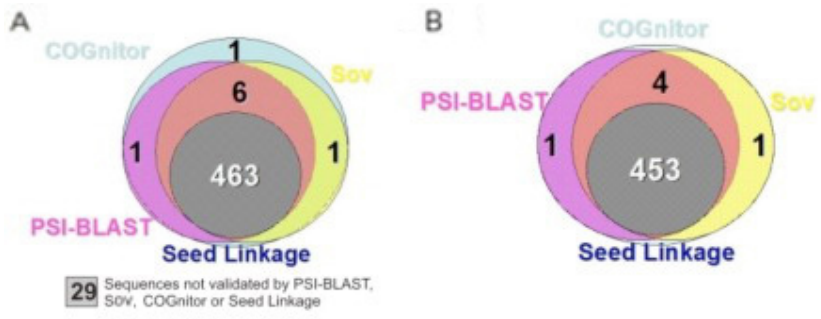

Figure 6. Diagrams combining four procedures of validation of UECOG0151. **A.** Five hundred and one recruited candidates were analyzed and 29 were not validated by any procedure. **B.** Same analysis but after 10% size selection filtering.

If a size filter of 10% is applied (see Material and Methods, and below), none of the 29 non-validated (false positive) entries were recruited. However, this filter also caused the exclusion of 10 of the 4-fold validated sequences (of 463), 6 of the PSI-BLAST and SOV validated sequences, plus the single candidates validated only by PSI-BLAST or only by SOV. To prevent these 18 candidates from being filtered out, the size selection limit would need to be increased. However, doing so might risk incorporating false positives. Figure 7 shows the fraction of all recruited candidates that are not incorporated as a function of the divergence of size selection allowed. Using a 10% cutoff, less than 10% of candidates are not recruited; the curve tends to saturate around 30%. However, using 30% would be acceptable only if coupling filtering with the validation procedures investigated here.

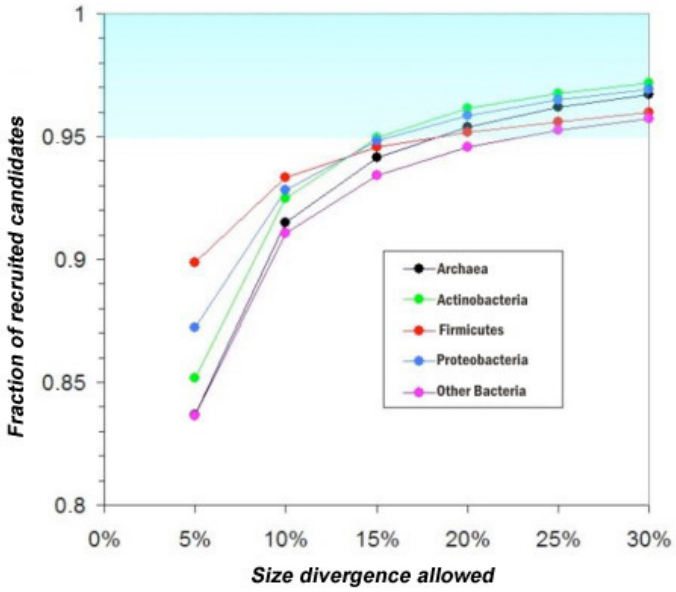

Figure 7. Fraction of recruitment as a function of size selection divergence. The experiment was conducted for the clades indicated.

In addition to the validation methods described above, we mapped domains using RPS-BLAST and the CDD database to determine if the domains shared by the recruiters for a given UniRef50 cluster are present also in the recruited candidates from that same cluster. The results are shown in Table 3. This procedure is stringent but nonetheless returned high values except for Archaea (73%) and Proteobacteria (84%), both of which fell below the statistically expected range; the results for other bacteria (83%) are in accordance with the results above.

Table 3. Validation of UECOG0151 with shared domains.

Clade	Recruiter	With domain	(%)	Recruited	Validated	(%)
Prokaryotes	51	38	75%	501	394	79%
Archaea	12	12	100%	30	22	73%
Bacteria	41	37	90%	471	393	83%
Actinobacteria	4	3	75%	41	39	95%
Firmicutes	8	7	87%	93	92	98%
Proteobacteria	21	20	95%	313	288	84%
Other bacteria	6	3	50%	24	20	83%

Finally, we examined whether recruited candidates are brothers in neighbor-joining trees. This experiment focused on Actinobacteria, Firmicutes and Proteobacteria since the inclusion of all clades could compromise the resolution in the tree. The entire set of sequences of each clade from UECOG0151 was used to construct neighbor-joining trees in five experiments and each recruiter was inspected to verify if the neighbor in the tree was from the same UniRef50 or was an original COG member. In the case that the neighbor was an original member, the algorithm continued the search without incrementing the score; the score was incremented only when the neighbor was a recruited member of the same UniRef50 cluster as the recruiter. The results shown in Table 4 together with manual inspection suggest that UniRef50 clusters are merged in neighbor-joining trees. Thus, the procedure does not seem to be appropriate for validation, but illustrates the important contribution of the enrichment for a better phylogenetic analysis of COGs.

Table 4. Analysis of UECOG0151 recruitment in neighbor-joining trees.

Cluster	Neighbor-joining validation						
	Recruiter	Recruited	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
Actinobacteria							
UniRef50_P65894	4	36	11	11	11	10	11
Firmicutes							
UniRef50_Q9HUV8	4	50	52	35	49	45	52
UniRef50_Q9ZF44	1	2	0	0	0	0	0
UniRef50_O66949	1	36	0	0	0	0	0
UniRef50_Q8K8Y4	1	19	19	19	19	19	19
UniRef50_Q5HH10	1	15	15	15	15	15	15
Sub-total	8	122	86	69	83	79	86
Proteobacteria							
UniRef50_O25817	2	2	2	2	2	2	2
UniRef50_Q8KBV8	2	67	8	8	8	4	3
UniRef50_Q9ABD2	1	8	1	1	1	2	1
UniRef50_Q9HUV8	1	9	4	4	5	6	6
UniRef50_Q9PC09	14	207	88	86	86	96	89
UniRef50_Q9PN47	1	18	16	16	16	13	16
Sub-total	21	311	119	117	118	123	117
Total	33	469	216	197	212	212	214

Reassessing a genome with UECOG procedure

The enrichment procedure was expected to add, in an identity range of 50% around each recruiter, a group of proteins from the respective UniRef50 cluster. Therefore, several proteins from organisms not included in COG would not be recruited unless a recruiter exists in a reasonable identity range. To illustrate this problem, we artificially deleted some genomes from COG and asked how many sequences of it would be recovered from other recruiters within COG (Edited COG). The results are presented in Table 5. While for *Escherichia coli* (txid 83333) the recovery reached 82% of all proteins present in UECOG, for some organisms this yield was not satisfactory, probably due to the lack of an organism closely related to this one in COG. Thus, the BBH procedure conducted with complete genomes seems to be very broad in comparison to the coverage attained by UniRef50 clusters, although the enrichment of the COG database of sequences closely related to the ones present in the database is significant.

Table 5. Recovery of a genome deleted from COG by recruiters in Edited COG.

Genome	Proteins present in			Recovered	(%)
	COG	Edited COG	UECOG		
Escherichia coli K12	3762	3242	3580	2938	82%
Bacillus halodurans	3262	3089	3182	1216	38%
Corynebacterium glutamicum	2249	2146	2693	531	20%
Archaeoglobus fulgidus	2034	1917	1920	378	20%

DISCUSSION

Databases of related proteins are a useful source for bioinformatics research, including annotation of novel genomes and phylogenetic studies. However, some approaches to generate such databases are limited by the need for complete genomes. One example is the COG database. Here we report a procedure to update and enrich the COG database to build UniRef-Enriched COGs, which will be maintained and made available at our website (http://biodados.icb.ufmg.br/uecog). The procedure consists of recruiting non-COG members of UniRef50 clusters that share one or more members with COGs. Only candidates that are not fragments are allowed to be incorporated. For this filtering, a parser of the UniProt FASTA files was necessary, but the current release of UniRef now contains a file with this information. We then took the opportunity to remove from Edited COGs the sequences that UniProtKB access labeled as a fragment. Edited COGs were obtained by updating entries that had new *gi* identifiers and by deleting entries that could not be reliably updated. Thus, UECOGs represent Edited COGs plus enrichment. A second important filter was to limit the enrichment to the clades present in the COG database. This procedure ensures enrichment with closely related sequences and avoids recruiting sequences from organisms that are too far apart from the ones possessing complete genomes. Using this approach, we safely obtained more data for analysis.

We also developed a series of approaches to validate the recruitment and illustrated their usage with the analysis of a chosen cluster UECOG0151. The approaches validating the highest percentage of sequences were PSI-BLAST and SOV validation, followed by Seed Link-

age software. COGnitor, available as a service in NCBI for single sequences, confirmed that 29 of the 30 sequences not validated by the three mentioned approaches were from a different cluster, COG0041, and suggested the use of a size filter to prevent false positives. Inspection of UECOG database indicates that it is feasible to use the validation procedures to restrict the inclusion of recruited candidates; inclusion of such validation steps is being considered for a second release of UECOG. The version under development will allow users to download only recruited candidates verified by the validation approaches described here. However, size selection was efficient for elimination of all recruited candidates that were not capable of validation. Further tests for the size selection are warranted, but there is indication that the use of 30% (rather than the 10% used in this construction), coupled with the validation procedures, could yield additional recruited candidates without diminishing the accuracy of recruitment. However, the routine use of such validations would compromise the speed of the update, and therefore was not considered an integral part of a generally applicable updated methodology.

One limitation of the approach described - possibly unique to COGs - is that it will fail to find remote orthologs; that is, it is limited to finding new members whose sequence identity is greater than or equal to 50% of any current member. COGs are noted for being independent of any similarity-based cutoff, and thus, the prototypical COG recruitment procedure (COGnitor) is able to recruit quite distant proteins. To estimate the contribution of low-identity orthologs, we examined the status of proteins in the current COG set. Specifically, we determined the number of proteins whose BLAST best hit was <50% identical. Of the 144,320 prokaryotic proteins, 52,608 (36.5%) would be missed by the UniRef50 recruitment procedure, and 95% of COGs would lack at least one member. The impact of the large number of false negatives largely depends on the purpose for which COGs can be used. On the one hand, COGs can be used for propagating annotation from one protein to another. For this purpose, the only real need is to have the two proteins in the same family. The impact of failing to cluster protein Y with protein X (when they should be together) is that annotation would not be propagated to protein Y (assuming protein X has some known function), and thus, protein Y would keep its current annotation. If protein Y is annotated and all the members of the COG that contain protein X are not, then these proteins would not get the annotation of Y, and again they would keep the current annotation. Accordingly, the failure to recruit all possible members does not really create a scientific problem above and beyond what already exists. Another purpose for which COGs can be used is to examine the metabolic suite of a given organism. For this purpose, it is imperative that all proteins from the organism are properly classified, lest one mistakenly concludes that a given set of proteins is missing from the organism. For this reason, UECOGs should not be used as is for such studies. However, it is expected that the grouping of more genomes in novel COG versions, or in any equivalent database built with the use of BBH relationships and complete genomes, can add important recruiters to positions in the evolutionary tree that could then recruit the missing proteins.

We envision that the UniRef50 enrichment procedure itself is applicable to any protein classification database, and may indeed perform quite well in certain systems. The advantage of the method is that it is not computationally intensive. This offers flexibility: use as is to enrich a database, perhaps for interim releases or for a quick update before annotation of new sequences, or couple the UniRef50 enrichment procedure with any flavored clustering methodology to reduce computation time and resources. No matter which database one wishes to update, the UniRef-based enrichment offers speed and accuracy.

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

Barbosa-Silva A, Satagopam VP, Schneider R and Ortega JM (2008). Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence. *BMC Bioinformatics* 9: 141.

Camon E, Magrane M, Barrell D, Lee V, et al. (2004). The gene ontology annotation (GOA) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Res.* 32: D262-D266.

Chen F, Mackey AJ, Stoeckert CJ Jr and Roos DS (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 34: D363-D368.

Chiu JC, Lee EK, Egan MG, Sarkar IN, et al. (2006). OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics* 22: 699-707.

Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, et al. (2006). Pfam: clans, web tools and services. *Nucleic Acids Res.* 34: D247-D251.

Fulton DL, Li YY, Laird MR, Horsman BG, et al. (2006). Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics* 7: 270.

Geourjon C, Combet C, Blanchet C and Deléage G (2001). Identification of related proteins with weak sequence identity using secondary structure information. *Protein Sci.* 10: 788-797.

Huang H, Barker WC, Chen Y and Wu CH (2003). iProClass: an integrated database of protein family, function and structure information. *Nucleic Acids Res.* 31: 390-392.

Krishnamurthy N, Brown D and Sjölander K (2007). FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. *BMC Evol. Biol.* 7 (Suppl 1): S12.

Li W and Godzik A (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658-1659.

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, et al. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* 30: 281-283.

O'Brien KP, Remm M and Sonnhammer EL (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* 33: D476-D480.

Rost B, Sander C and Schneider R (1994). Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* 235: 13-26.

Schultz J, Milpetz F, Bork P and Ponting CP (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U.S.A.* 95: 5857-5864.

Suzek BE, Huang H, McGarvey P, Mazumder R, et al. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23: 1282-1288.

Tatusov RL, Koonin EV and Lipman DJ (1997). A genomic perspective on protein families. *Science* 278: 631-637.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41.

UniProt Consortium (2007). The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 35: D193-D197.

Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Testing the performance of automated annotation of ESTs with the Kegg Orthology (KO) database demonstrates lack of completeness of clusters

G.R. Fernandes, M.A. Mudado and J.M. Ortega

Laboratório de Biodados, Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil

Corresponding author: J.M. Ortega
E-mail: miguel@icb.ufmg.br

Genet. Mol. Res. 7 (3): 948-957 (2008)
Received June 2, 2008
Accepted august 11, 2008
Published September 30, 2008

ABSTRACT. The KEGG Orthology (KO) database was tested as a source for automated annotation of expressed sequence tags (ESTs). We used a control experiment where every EST was assigned to its cognate protein, and an annotation experiment where the ESTs were annotated by proteins from other organisms. Analyzing the results, we could assign classes to the annotation: correct, changed and speculated. The correct annotation ranged from 57 (*Caenorhabditis elegans*) to 81% (*Homo sapiens*). In spite of the changed annotation being low (1 in *H. sapiens* to 9% in *Arabidopsis thaliana*), the speculation was very high (18 in *H. sapiens* to 38% in *C. elegans*). We propose eliminating part of the speculated annotation using the KEGG Genes database to enrich KO clusters, decreasing the speculation from 38 to 2% in *C. elegans*. Thus, the KO database still demands some effort for moving sequences from Kegg GENES to KO, to complement the annotation performance.

Key words: KEGG Orthology; Annotation; Orthologs; BLAST; Expressed sequence tags

INTRODUCTION

Ever since the first complete genome of a cellular life form was described in 1995, the analyses and identification of genes and their function have become a challenge (Fleischmann et al., 1995; Brosius, 1996). Using computers to analyze the information from a well-understood organism, it is possible to transfer this knowledge to poorly characterized genomes, based on the similarity shared by their DNA or protein sequences (Tatusov et al., 1996). These similarity searches are usually conducted with the usage of several softwares available in the BLAST package (http://www.ncbi.nlm.nih.gov/BLAST/) (Altschul et al., 1997). A relationship of homology can be established by selecting the higher bit score and using an E-value cutoff requirement (Koonin and Galperin, 2003; Koonin et al., 2004). The quality of this type of annotation depends considerably on the quality, reliability and completeness of the information stored in the target database for BLAST alignment (Mudado et al., 2005). Secondary databases have been used as a source of information to annotate novel genome and transcriptome projects (Vettore et al., 2003). These databases usually contain genomics and proteomics data; furthermore, they provide function, structure, and other complementary information.

One of the best known secondary databases is KEGG (Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg). KEGG is a web resource directed at integrating genomic and functional information, and has been used to standardize gene annotation (Kanehisa and Goto, 2000). Up to November 2006, the KEGG database was composed of 1,775,476 gene entries derived from the genomes of 446 organisms (including 35 eukaryotes, 382 bacteria and 29 archea). The KEGG database information is organized into four sub-databases: GENES, BRITE, PATHWAY and LIGAND. GENES contains gene catalogs of completely sequenced genomes and some partial genomes. BRITE informs about protein-protein interactions and relationships. The PATHWAY database informs about generalized protein interaction networks (pathway and complexes), where several cellular processes are involved. LIGAND informs about chemical compounds and chemical reactions that are relevant to cellular processes (Kanehisa et al., 2002). KEGG Orthology (KO) was developed to integrate pathway and genomic information in KEGG. KO was introduced to replace the Enzyme Commission (EC) number as identifier of gene product in metabolic pathways. In order to classify gene functions, KO is based on computational analyses and manual curation of the SSDB (Sequence Similarity Database) ortholog clusters. KO is structured as a hierarchy of four flat levels, and this structure allows KO to be a great putative source for automated annotation (Mao et al., 2005).

In the present study, we tested the potential of the KO database to automatically annotate ESTs of model organisms using BLAST, with a methodology we proposed earlier (Mudado et al., 2005). Although the annotation of expressed sequence tags (EST) with KO is a common procedure, the tests of performance conducted here are informative and aimed at addressing the confidence of the method. We first assigned the ESTs of a given organism to their proteins deposited in KO. Afterwards, we used the remaining proteins from the other organisms in KO, without the proteins of the cognate organism used in the assignment, to annotate the previously assigned ESTs, evaluating the correctness of the results. The resulting data show that the annotation accuracy was considerably high. However, in spite of the changed annotation being low (1 in *Homus sapiens* to 9% in *Arabidopsis thaliana*), the speculated annotation (defined here as annotation in absence of assignment) was unexpectedly very high (18 in *H. sapiens* to 38% in *Caenorhabditis elegans*) as compared to results previously obtained with the KOG database (Mudado et al., 2005). We

showed evidence that many ESTs that receive speculated annotation can be assigned to protein entries in KEGG GENES. We propose eliminating part of the speculated annotation using the KEGG GENES database to enrich KO clusters, which markedly decreases the speculation from 38 to 2% in *C. elegans*. Thus, the KO database still demands some effort for moving sequences from KEGG GENES to KO, to complement the annotation performance.

MATERIAL AND METHODS

Download and selection of ESTs

ESTs were retrieved from the NCBI web site (www://ncbi.nlm.nih.gov). We downloaded collections of ESTs from four model organisms: *A. thaliana*, *C. elegans*, *Drosophila melanogaster*, and *H. sapiens*. We chose to download GenBank flat files and filter these files with a Perl script to select ESTs regarding information about, for example, organ, tissue, library, author, development stage, and sequence length. These data were used to populate an MySQL database, which was used to retrieve ESTs from healthy tissues and from libraries with more than five thousand entries (see Table 1).

Proteins

Protein sequences were downloaded from the KEGG web site (ftp://ftp.genome.jp/pub/kegg/). We used sequences from ten organisms in the analysis: *A. thaliana* (Ath), *C. elegans* (Cel), *Canis familiaris* (Cfa), *D. melanogaster* (Dme), *Encephalitozoon cuniculi* (Ecu), *H. sapiens* (Hsa), *Mus musculus* (Mmu), *Rattus norvegicus* (Rno), *Schizosaccharomyces pombe* (Spo), and *Saccharomyces cerevisiae* (Sce). Only proteins present in the KO database were used, which represent a small proportion of the total available proteins from KEGG GENES, as shown in Figure 1.

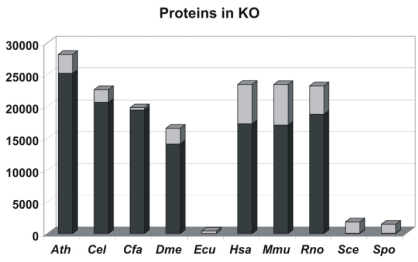

Figure 1. Distribution of the organism's proteins into KEGG Orthology (KO) clusters. The light grey bars represent the proteins within KO, while dark grey bars those outside KO.

BLAST experiments

The BLAST experiment was divided into two stages (Figure 2): initially, we executed a tBLASTn using the KO proteins from one organism against its own EST repository (e.g., Ath

proteins *vs* Ath ESTs). This stage works as a positive control, which assigns a given EST to its cognate protein in the database. The low complexity filter was disabled and an E-value cutoff of at least 10^{-10} was established in order to favor alignments of only homologous sequences. Additionally, an assignment identity cutoff was applied only for this step. Assignment cutoffs were determined as described by Mudado et al. (2005 and Mudado MA, Fernandes GR and Ortega JM, unpublished results). In the second step, we aligned the ESTs with the database depleted of the cognate organism proteins. We used the same threshold for low complexity filter and E-value established above. This experiment was aimed at simulating the annotation procedure of a novel transcriptome, as the ESTs were aligned only with proteins from organisms different from its source organism. A summary of both stages is shown in Figure 2. This procedure has already been performed with the KOG database (Mudado et al., 2005) in order to evaluate the database potential of annotating.

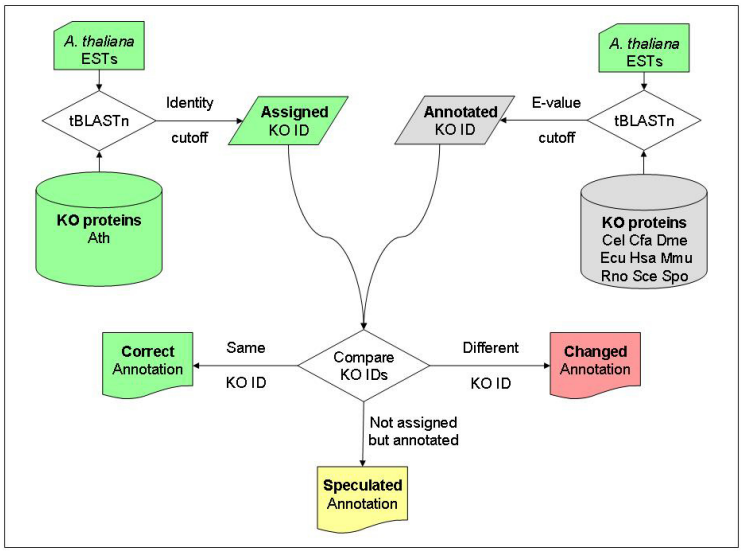

Figure 2. Representation of the performance tests using expressed sequence tags (ESTs) from *Arabidopsis thaliana*. On the left, alignments requiring an identity cutoff (78% for *A. thaliana*) on top of a 10^{-10} E-value cutoff (data not shown) serve as a positive control, and assignment of a KEGG Orthology (KO) identifier (KO ID) to each EST. On the right, a second step simulates a novel genome annotation. Comparing the results, we labeled the annotation as correct, changed or speculated.

Assignment cutoff

Briefly (see Mudado et al., 2005, for details), 50 proteins without paralogs with the highest expression determined by the number of hits to EST were selected for analysis. ESTs were aligned to either the nucleotide sequence of those proteins' CDS with BLASTn or to those proteins with tBLASTn, under a stringent E-value cutoff of 1e-10 (note that to preserve the E-value determination the formatted database used in both steps was the total set of ESTs).

We first determined the percentage of aligned ESTs that show over 96% identity at the nucleotide to nucleotide level. We then determined the identity cutoff at the nucleotide to amino acid level, which groups this same percentage of hits, and used this value as identity cutoff for EST assignment to the correct protein.

Annotation classes

Classes of annotation were established by comparing the BLAST results of the stages 1 (assignment) and 2 (annotation). ESTs within the correct annotation category were assigned and annotated by the same KO protein from stages 1 and 2. Changed annotation category has ESTs with different assigned and annotated KO proteins. Speculated annotation occurs when an EST is annotated by a KO protein but it has not been assigned by any protein (so we have no positive control for that annotation).

KEGG GENES

The KEGG GENES database was also used to assign ESTs in order to try to minimize speculated annotation. This database involves all proteins from an organism that is present in KEGG.

Accuracy

The accuracy was measured by the PPV (positive predictive value). In this case the PPV can be defined as the quotient of the correct annotation divided by the correct and changed annotations.

RESULTS AND DISCUSSION

Annotation overview

After downloading and filtering all ESTs (Table 1), the BLAST searches were performed and the annotation divided into categories (see Material and Methods).

Table 1. Total number of expressed sequence tag (EST) downloaded and total remaining after MySQL filtering.

Organism	EST - Downloaded	EST - Used
Arabidopsis thaliana	622,788	360,833
Caenorhabditis elegans	302,080	293,530
Drosophila melanogaster	383,407	375,360
Homo sapiens	1,673,145	365,619

With the default identity cutoffs (78, 81, 71, and 72%), annotation was distributed as follows: 71, 57, 70, and 81% of correct annotation; 9, 5, 5, and 1% of changed annotation, and 20, 38, 25, and 18% of speculated annotation, for Ath, Cel, Dme, and Hsa, respectively. Although changed annotation was low, the database provided very high levels of speculated annotation, suggesting lack of completeness of the KO clusters.

The total annotation covered 21, 36, 32.5, and 37.1% of all ESTs from Ath, Cel, Dme, and Hsa, respectively. By imposing an identity cutoff, the proportion of correct annotations decreased and the proportion of speculated annotation increased until it reaches a less variable level if the assignment cutoff is between 70-80% identity (Figure 3; see also discussion about accuracy below). This effect is mostly caused by a decrease in EST assignment if the cutoff is too stringent in the first stage of the procedure (Figure 2). However, the changed annotation does not show apparent changes as the cutoff varies. Based on previous results with the KOG database, we believe that changed annotation is mainly caused by proteins that are specific to an organism but can be annotated by similar proteins, which share the same conserved domains. With the KOG database (Mudado et al., 2005), a clear plateau is observed with low values of assignment cutoffs, and that is possibly explained by the more completeness of this database as compared to KO (see below). Although the cutoff experimentally defined by us has experimental support, the results were obtained with all possible cutoffs for comparison, and the conclusion is that the correct and changed annotations both do not vary so much if a lower cutoff is chosen.

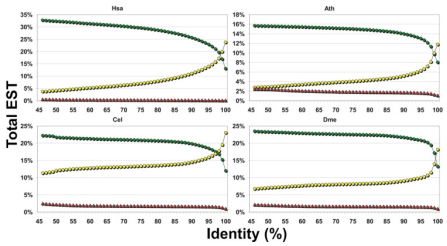

Figure 3. Proportion of the annotation classes and their relationship with the identity cutoff. The green circles represent the correct annotation, the yellow circles are the speculative annotation, and the red triangles represent the changed annotation. EST = expressed sequence tag.

Coverage

The KO coverage (percentage of KO clusters that find a hit in a representative collection of ESTs) is an important aspect to be analyzed, since it tells how relevant the KO groups are for the annotation of reasonably large collections of ESTs. The coverage overview is shown in Figure 4. KO includes several organisms and a large set was used in this study (Figure 1). Taking into consideration the KO groups of all these organisms (upper panels), it was observed that the coverage during the annotation (referred to here as "Sampling" of the EST collection with KO clusters, panels on the left) was higher than during the assignment (referred to here as "Picturing" the EST content supposedly assigned to the correct protein, panels on the right). Hsa clusters apparently are the most complete. Accordingly, this effect is reduced when only KO clusters that contain proteins of the studied organism are used (lower panels). Here, Sampling yields almost the same coverage as Picturing (even for Ath, the only plant), indicating that similarity between cluster members is sufficient. Thus, coverage does not seem to be substantially affected by the lack of paralogs in clusters that contain at least a representative of the organism whose ESTs are being analyzed.

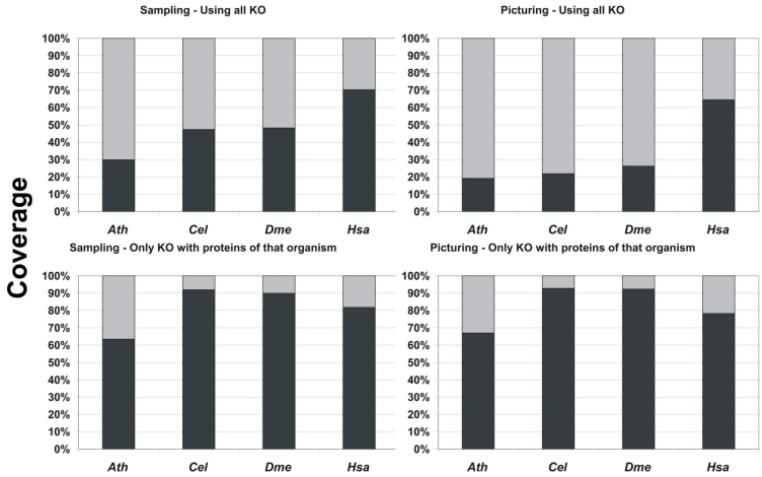

Figure 4. Percentage of KEGG Orthology (KO) clusters used for the expressed sequence tags (EST) assignment (picturing) and annotation (sampling) of Ath, Cel, Dme, and Hsa. In the two upper graphics, we used the KO clusters of all available organisms, while in the graphics at the bottom, we used as reference only the KO groups, which contain proteins of the same organism to which the EST belongs. The dark bars indicate the KO groups that had a protein that matched with an EST, and the light bars represent the cluster that had no matches with this EST collection.

Extra-KO annotation

The KEGG GENES database was complementarily used in order to try to minimize speculated annotation. This database contains all identified proteins of an organism present in KEGG. A graphical representation of this search is shown in Figure 5.

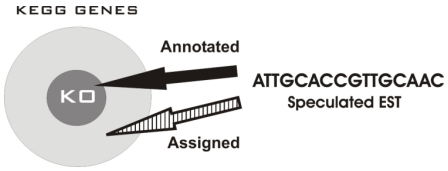

Figure 5. Procedure of the extra-KO annotation. An expressed sequence tag (EST) that had a speculated annotation was submitted to a BLAST against all KEGG GENES proteins of the same EST's organism. We followed the same parameters of E-value cutoff and low complexity filter used in the previous BLAST experiment. The same identity cutoff was applied to determine the assignment. KO = KEGG Orthology.

Some examples of KO clusters related with speculated annotation, and their respective extra-KO matches are shown in Table 2. The first column shows the organism to which the EST belongs. The second column represents the cluster entry that contains the protein, which matched the EST leading to the speculated annotation. The third shows how many times that KO cluster was related to a speculation. The fourth and fifth columns show the KO cluster name and

the extra-KO protein name, respectively. The data suggest that speculated annotation could turn into correct annotation as the KO clusters are enlarged to group entries from KEGG GENES.

Table 2. Relationship of speculation and extra-KO assignment.

Organism	KO	Amount	Cluster name	Extra-KO protein name
Ath	K05692	805	Actin, beta/gamma, cytoplasmic	F27J15.1; actin 8 (ACT8)
Cel	K01829	2622	Protein disulfide-isomerase	Protein disulfide isomerase protein 2, isoform a
Dme	K04439	1699	Arrestin	Arrestin

KO = KEGG Orthology.

Considering the cutoff values of 78, 81, 71, and 72% for Ath, Cel, Dme, and Hsa, respectively, the distribution of the annotation classes after the "extra-KO" experiment is shown in Figure 6. Note that the light green area would consist of speculated annotation (yellow) if the KEGG GENES assignment was not considered. This area is highlighted in light green in supposition that its candidates can be turned into correct annotation.

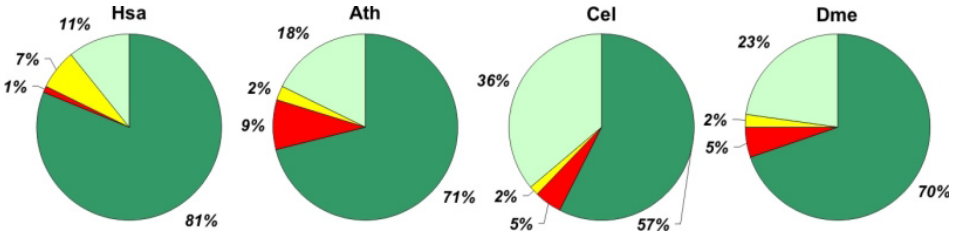

Figure 6. Pie graphs indicating the distribution of the annotation classes with an identity cutoff value of 72, 78, 81, and 71% for Hsa, Ath, Cel, and Dme, respectively. The dark green area represents the correct annotation, the light green represents the EST with speculated annotation, which has matches in the extra-KO proteins, the yellow area represents the speculated annotation that was not solved by the extra-KO assignment, and the red area indicates the changed annotation proportion.

Since *C. elegans* was the organism demonstrating the highest proportion of speculated annotation, we decided to evaluate a group of biochemical pathways involved in amino acid metabolism to illustrate this point. Table 3 lists sixteen pathways as obtained from the KO database and the number of KO entries in them, adding up to 692 KO clusters. Note that some pathways either consist of or include non-essential amino acid metabolism, and thus, *C. elegans* may lack proteins in some clusters such as methionine metabolism. However, we detected speculated annotation of ESTs to those clusters in 186 events. In a total of 152 cases, we were able to assign those ESTs to proteins from the KEGG GENES database that were not included in KO, indicating that these proteins are prompted to be included in the KO clusters. Moreover, when sequences from certain KO pathways are selected to evaluate expression in other organisms such as worms, the analysis could be compromised by the lack of entries in the KO database. We believe that the high levels of speculated annotation observed are probably due to a cautious but incomplete evolution of KO clusters as compared to the KOG database.

Table 3. Number of KEGG Orthology (KO) clusters from amino acid metabolism pathways that yield speculated annotation of *Caenorhabditis elegans* expressed sequence tags (ESTs) but with ESTs assigned to Kegg GENES extra-KO *C. elegans* proteins.

Pathway*	Number of KO in pathway	Lacking cel proteins	With speculated annotation	With extra-KO assignment
Total	692	511	186	152 (22%)

*Consist of or include biosynthesis pathway of non-essential amino acids.

Accuracy

The annotation method accuracy for each organism is shown in Figure 7, where the curves represent the PPV value fluctuation for each organism. The best accuracy was obtained for the Hsa annotation. It can be explained by the fact that the KO database uses this organism as reference for the ortholog groups. The worst accuracy, but acceptable, was obtained for Ath. This might have occurred because it is the only plant in the database.

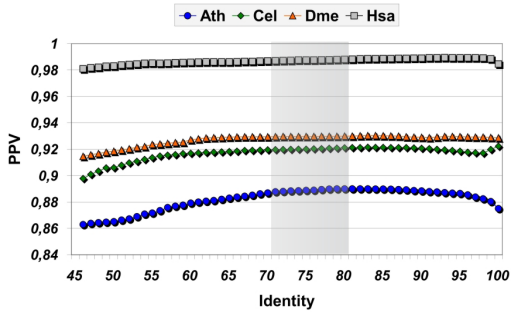

Figure 7. Annotation accuracy fluctuation related to identity among the sequences. Note: identity in percentages, and positive predictive value (PPV) varying from 0-1.

CONCLUSION

We demonstrated that, although being an incomplete database, KO is a good vocabulary source for automated annotation. This was proved by the high accuracy when assigned and annotated datasets were compared among the distinct organisms that build KO. For this purpose, it is also crucial that identity cutoff be selected in such a way that a great coverage of correctly annotated ESTs balances with a low proportion of changed annotation. The PPV test gave us an overview of this value, which allowed us to confirm the appropriated cutoff values. Additionally, the lack of entries that are present in KEGG GENES can explain the high amount of speculated annotation, suggesting that additional evolution of the KO database with inclusion of entries in KEGG GENES will greatly enhance the performance of automated annotation with KO. This could be seen in the extra-KO experiment, where proteins without a KO cluster were used to minimize the speculation. In summary, automated annotation with KO is accurate, and the analysis presented here supports the inclusion of KEGG GENES entries in KO clusters for enhanced performance of annotation with the KO database.

ACKNOWLEDGMENTS

We thank Dr. Darren Natale from PIR for critically reviewing this manuscript. G.R. Fernandes and M.A. Mudado were recipients of fellowships from CAPES. Laboratório de Biodados had a grant from FAPEMIG as a member of Minas Gerais Genome Network.

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

Brosius J (1996). More *Haemophilus* and *Mycoplasma* genes. *Science* 271: 1302-1304.

Fleischmann RD, Adams MD, White O, Clayton RA, et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496-512.

Kanehisa M and Goto S (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28: 27-30.

Kanehisa M, Goto S, Kawashima S and Nakaya A (2002). The KEGG databases at GenomeNet. *Nucleic Acids Res.* 30: 42-46.

Koonin EV and Galperin MY (2003). Sequence-Evolution-Function. Computational Approaches in Comparative Genomics. Kluwer Academic Publishers, Norwell.

Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, et al. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* 5: R7.

Mao X, Cai T, Olyarchuk JG and Wei L (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* 21: 3787-3793.

Mudado MA, Bravo-Neto E and Ortega JM (2005). Tests of automatic annotation using KOG proteins and ESTs from 4 eukaryotic organisms. *Lecture Notes Computer Sci.* 3594: 141-152.

Tatusov RL, Mushegian AR, Bork P, Brown NP, et al. (1996). Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.* 6: 279-291.

Vettore AL, da Silva FR, Kemper EL, Souza GM, et al. (2003). Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Res.* 13: 2725-2735.

ARTICLE

doi:10.1038/nature09944

Enterotypes of the human gut microbiome

Manimozhiyan Arumugam[1]*, Jeroen Raes[1,2]*, Eric Pelletier[3,4,5], Denis Le Paslier[3,4,5], Takuji Yamada[1], Daniel R. Mende[1], Gabriel R. Fernandes[1,6], Julien Tap[1,7], Thomas Bruls[3,4,5], Jean-Michel Batto[7], Marcelo Bertalan[8], Natalia Borruel[9], Francesc Casellas[9], Leyden Fernandez[10], Laurent Gautier[8], Torben Hansen[11,12], Masahira Hattori[13], Tetsuya Hayashi[14], Michiel Kleerebezem[15], Ken Kurokawa[16], Marion Leclerc[7], Florence Levenez[7], Chaysavanh Manichanh[9], H. Bjørn Nielsen[8], Trine Nielsen[11], Nicolas Pons[7], Julie Poulain[3], Junjie Qin[17], Thomas Sicheritz-Ponten[8,18], Sebastian Tims[15], David Torrents[10,19], Edgardo Ugarte[3], Erwin G. Zoetendal[15], Jun Wang[17,20], Francisco Guarner[9], Oluf Pedersen[11,21,22,23], Willem M. de Vos[15,24], Søren Brunak[8], Joel Doré[7], MetaHIT Consortium†, Jean Weissenbach[3,4,5], S. Dusko Ehrlich[7] & Peer Bork[1,25]

Our knowledge of species and functional composition of the human gut microbiome is rapidly increasing, but it is still based on very few cohorts and little is known about variation across the world. By combining 22 newly sequenced faecal metagenomes of individuals from four countries with previously published data sets, here we identify three robust clusters (referred to as enterotypes hereafter) that are not nation or continent specific. We also confirmed the enterotypes in two published, larger cohorts, indicating that intestinal microbiota variation is generally stratified, not continuous. This indicates further the existence of a limited number of well-balanced host–microbial symbiotic states that might respond differently to diet and drug intake. The enterotypes are mostly driven by species composition, but abundant molecular functions are not necessarily provided by abundant species, highlighting the importance of a functional analysis to understand microbial communities. Although individual host properties such as body mass index, age, or gender cannot explain the observed enterotypes, data-driven marker genes or functional modules can be identified for each of these host properties. For example, twelve genes significantly correlate with age and three functional modules with the body mass index, hinting at a diagnostic potential of microbial markers.

Various studies of the human intestinal tract microbiome based on the 16S ribosomal-RNA-encoding gene reported species diversity within and between individuals[1–3], and the first metagenomics studies characterized the functional repertoire of the microbiomes of several American[4,5] and Japanese[6] individuals. Although a general consensus about the phylum level composition in the human gut is emerging[1,3,7], the variation in species composition[1,2] and gene pools[5,8] within the human population is less clear. Furthermore, it is unknown whether inter-individual variation manifests itself as a continuum of different community compositions or whether individual gut microbiota congregate around preferred, balanced and stable community compositions that can be classified. Studying such questions is complicated by the complexity of sampling, DNA preparation, processing, sequencing and analysis protocols[9] as well as by varying physiological, nutritional and environmental conditions. To analyse the feasibility of comparative metagenomics of the human gut across cohorts and protocols and to obtain first insights into commonalities and differences between gut microbiomes across different populations, we Sanger-sequenced 22 European metagenomes from Danish, French, Italian and Spanish individuals that were selected for diversity (Supplementary Notes section 1), and combined them with existing Sanger (13 Japanese[6], 2 American[4]) and pyrosequencing (2 American[5]) gut data sets—totalling 39 individuals.

Global variation of human gut metagenomes

The vast majority of sequences in the newly sequenced 22 European samples belong to bacteria—only 0.14% of the reads could be classified as human contamination, all other eukaryotes together only comprised 0.5%, archaea 0.8% and viruses up to 5.8% (see Supplementary Notes section 2.1 for details).

To investigate the phylogenetic composition of the 39 samples from 6 nationalities, we mapped metagenomic reads, using DNA sequence homology, to 1,511 reference genomes (Supplementary Table 3) including 379 publicly available human microbiome genomes generated through the National Institutes of Health (NIH) Human Microbiome Project[10] and the European MetaHIT consortium[11] (Supplementary Methods section 4.1). To consistently estimate the functional composition of the samples, we annotated the predicted genes from the metagenomes using eggNOG[12] orthologous groups (Supplementary Methods section 6.2). We ensured that comparative analysis using these procedures was not biased by data-set origin, sample preparation, sequencing technology and quality filtering (see Supplementary Notes section 1).

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [2]VIB—Vrije Universiteit Brussel, 1050 Brussels, Belgium. [3]Commissariat à l'Energie Atomique, Genoscope, 91000 Evry, France. [4]Centre National de la Recherche Scientifique, UMR8030, 91000 Evry, France. [5]Université d'Evry Val d'Essone 91000 Evry, France. [6]Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, 31270-901 Belo Horizonte, Minas Gerais, Brazil. [7]Institut National de la Recherche Agronomique, 78350 Jouy en Josas, France. [8]Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Lyngby, Denmark. [9]Digestive System Research Unit, University Hospital Vall d'Hebron, Ciberehd, 08035 Barcelona, Spain. [10]Barcelona Supercomputing Center, Jordi Girona 31, 08034 Barcelona, Spain. [11]Marie Krogh Center for Metabolic Research, Section of Metabolic Genetics, Faculty of Health Sciences, University of Copenhagen, DK-2100 Copenhagen, Denmark. [12]Faculty of Health Sciences, University of Southern Denmark, DK-5000 Odense, Denmark. [13]Computational Biology Laboratory Bld, The University of Tokyo Kashiwa Campus, Kashiwa-no-ha 5-1-5, Kashiwa, Chiba, 277-8561, Japan. [14]Division of Bioenvironmental Science, Frontier Science Research Center, University of Miyazaki, 5200 Kiyotake, Miyazaki 889-1692, Japan. [15]Laboratory of Microbiology, Wageningen University, 6710BA Ede, The Netherlands. [16]Tokyo Institute of Technology, Graduate School of Bioscience and Biotechnology, Department of Biological Information, 4259 Nagatsuta-cho, Midori-ku, Yokohama-shi, Kanagawa Pref. 226-8501, Japan. [17]BGI-Shenzhen, Shenzhen 518083, China. [18]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2800 Lyngby, Denmark. [19]Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain. [20]Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark. [21]Institute of Biomedical Science, Faculty of Health Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark. [22]Hagedorn Research Institute, DK-2820 Gentofte, Denmark. [23]Faculty of Health Sciences, University of Aarhus, DK-8000 Aarhus, Denmark. [24]University of Helsinki, FI-00014 Helsinki, Finland. [25]Max Delbrück Centre for Molecular Medicine, D-13092 Berlin, Germany.

*These authors contributed equally to this work.

†Lists of authors and affiliations appear at the end of the paper.

We also investigated whether the relatively low and somewhat arbitrary amounts of sequence per sample (between 53–295 Mb) bias our results: we assigned habitat information to 1,368 of the 1,511 reference genomes, distinguished between orthologous groups from gut and 'non-gut' species and conclude that our data set captures most of the functions from gut species even though functions from non-gut species accumulated with each additional sample (Fig. 1a; see Supplementary Notes section 1.3).

We then characterized the phylogenetic variation across samples at the genus and phylum levels, and functional variation at gene and functional class levels. As infants are known to have very heterogeneous, unstable and distinctive microbiota[6,13], we excluded the four respective Japanese samples from the analysis. Using calibrated similarity cutoffs (Supplementary Fig. 1), on average, 52.8% of the fragments in each sample could be robustly assigned to a genus in our reference genome set (ranging from 22% to 80.5%), and 80% could be assigned to a phylum (ranging from 64.9% to 91%) implying that the trends observed (Fig. 1b) represent a large fraction of the metagenome.

The phylogenetic composition of the newly sequenced samples confirms that the Firmicutes and Bacteroidetes phyla constitute the vast majority of the dominant human gut microbiota[7] (Fig. 1b, inset). *Bacteroides* was the most abundant but also most variable genus across samples (Fig. 1b and Supplementary Notes section 2.2), agreeing with previous observations[6,14]. Our function identification protocol led to a high functional assignment rate: 63.5% of all predicted genes in the Sanger-sequenced samples analysed (41% of all predicted genes in two samples obtained by pyrosequencing; Supplementary Table 5) can be assigned to orthologous groups, and orthologous group abundance patterns agree with previous observations[6,15] (for example, histidine kinases make up the largest group; Fig. 1c and Supplementary Notes section 2.3).

Abundant functions from low-abundance microbes

Microbes in the human gut undergo selective pressure from the host as well as from microbial competitors. This typically leads to a homeostasis of the ecosystem in which some species occur in high and many in low abundance[16] (the 'long-tail' effect, as seen in Fig. 1b), with some low-abundance species, like methanogens[17], performing specialized functions beneficial to the host. Metagenomics enables us to study the presence of abundant functions shared by several low-abundance

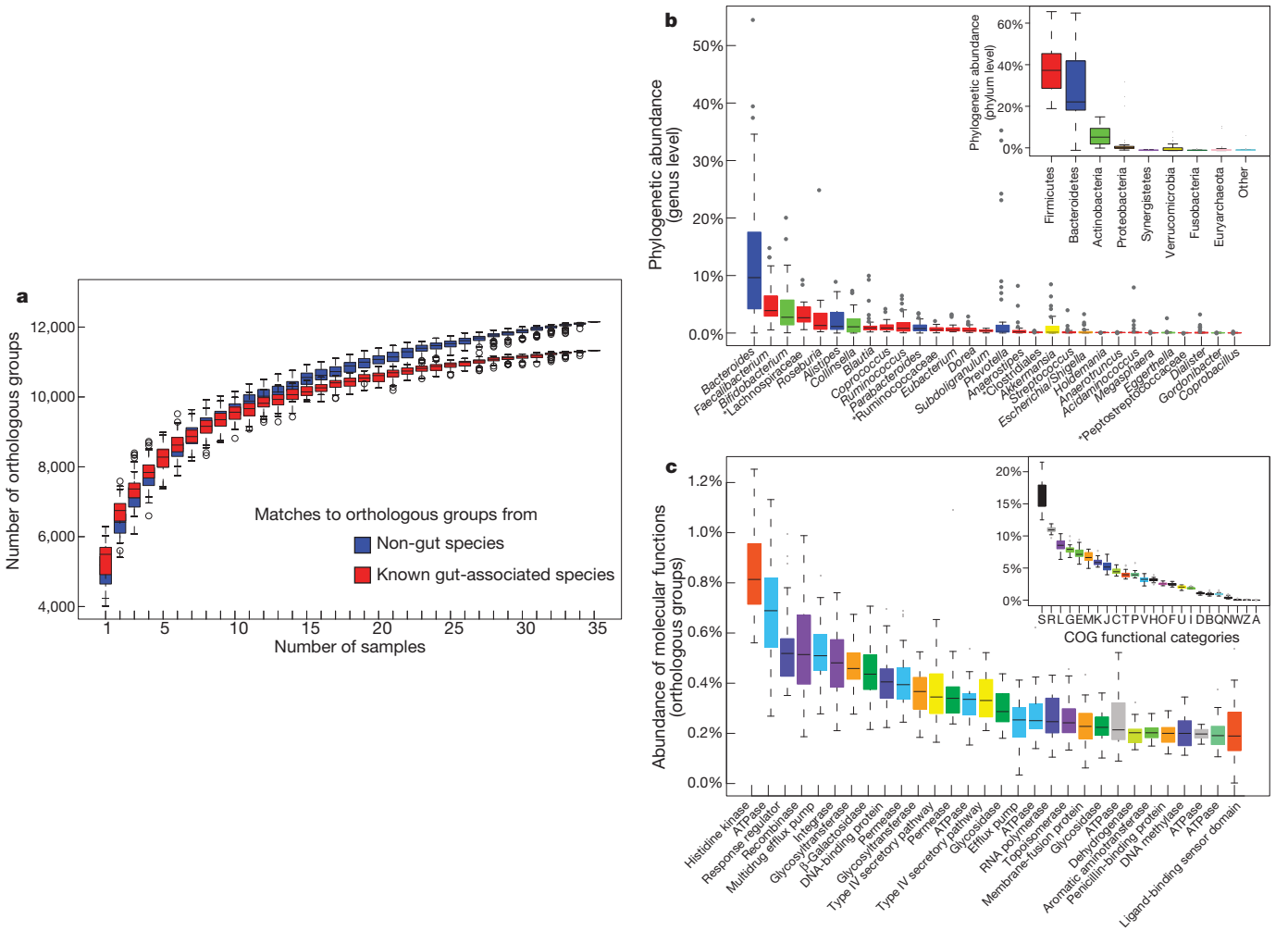

Figure 1 | Functional and phylogenetic profiles of human gut microbiome. **a**, Simulation of the detection of distinct orthologous groups when increasing the number of individuals (samples). Complete genomes were classified by habitat information and the orthologous groups divided into those that occur in known gut species (red) and those that have not yet been associated with gut (blue). The former are close to saturation when sampling 35 individuals (excluding infants) whereas functions from non-gut (probably rare and transient) species are not. **b**, Genus abundance variation box plot for the 30 most abundant genera as determined by read abundance. Genera are coloured by their respective phylum (see inset for colour key). Inset shows phylum abundance box plot. Genus and phylum level abundances were measured using reference-genome-based mapping with 85% and 65% sequence similarity cutoffs. Unclassified genera under a higher rank are marked by asterisks. **c**, Orthologous group abundance variation box plot for the 30 most abundant orthologous gruops as determined by assignment to eggNOG[12]. Orthologous groups are coloured by their respective functional category (see inset for colour key). Inset shows abundance box plot of 24 functional categories. Boxes represent the interquartile range (IQR) between first and third quartiles and the line inside represents the median. Whiskers denote the lowest and highest values within 1.5 × IQR from the first and third quartiles, respectively. Circles represent outliers beyond the whiskers.

species, which could shed light on their survival strategies in the human gut. In the samples analysed here, the most abundant molecular functions generally trace back to the most dominant species. However, we identified some abundant orthologous groups that are contributed to primarily by low-abundance genera (see Supplementary Fig. 2, Supplementary Table 6 and Supplementary Notes section 3). For example, low-abundance *Escherichia* contribute over 90% of two abundant proteins associated with bacterial pilus assembly, FimA (COG3539) and PapC (COG3188), found in one individual (IT-AD-5). Pili enable the microbes to colonize the epithelium of specific host organs; they help microbes to stay longer in the human intestinal tract by binding to human mucus or mannose sugars present on intestinal surface structures[18]. They are also key components in the transfer of plasmids between bacteria through conjugation, often leading to exchange of protective functions such as antibiotic resistance[18]. Pili can thus provide multiple benefits to these low-abundance microbes in their efforts

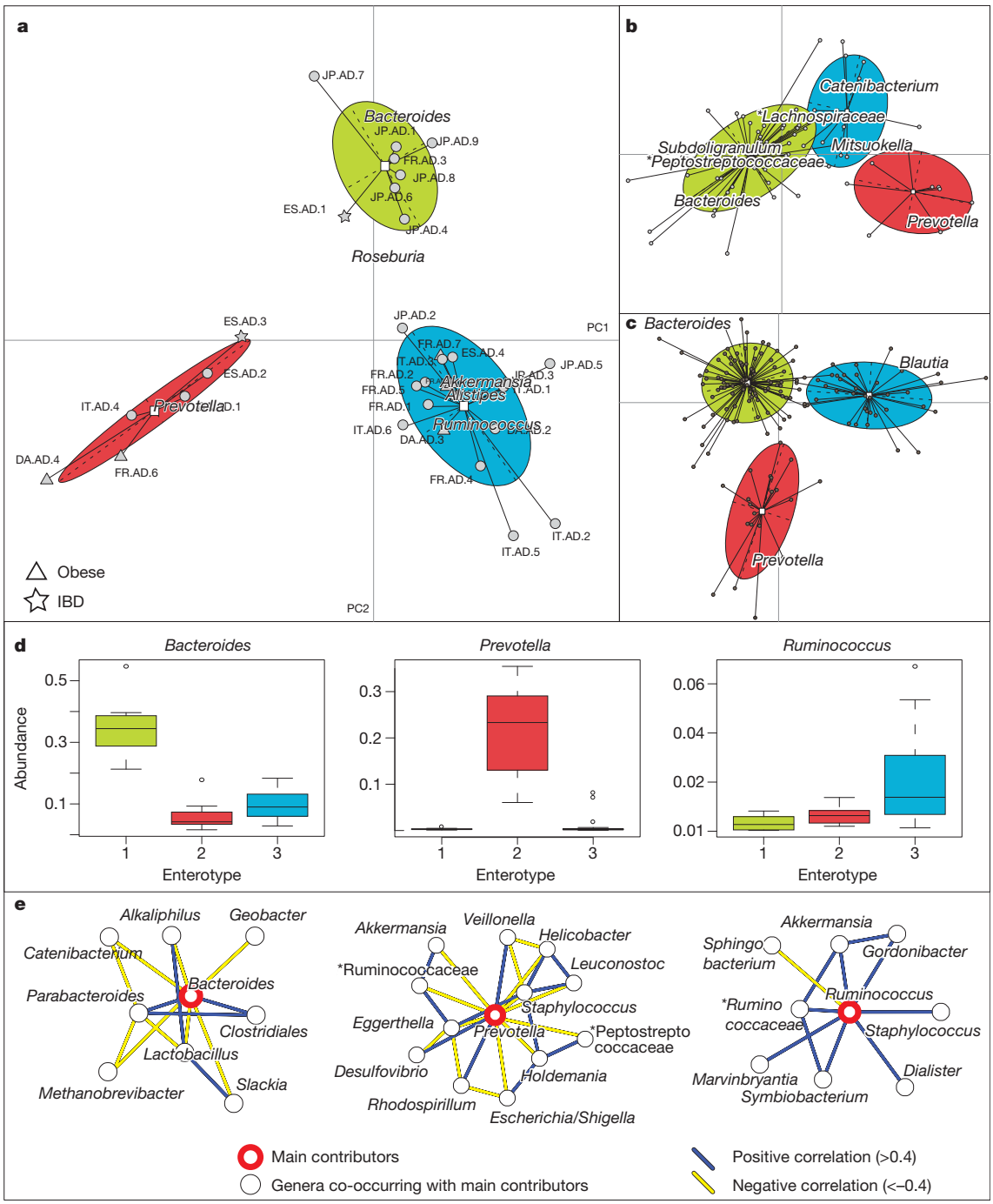

Figure 2 | Phylogenetic differences between enterotypes. a–c, Between-class analysis, which visualizes results from PCA and clustering, of the genus compositions of 33 Sanger metagenomes estimated by mapping the metagenome reads to 1,511 reference genome sequences using an 85% similarity threshold (**a**), Danish subset containing 85 metagenomes from a published Illumina data set[8] (**b**) and 154 pyrosequencing-based 16S sequences[5] (**c**) reveal three robust clusters that we call enterotypes. IBD, inflammatory bowel disease. Two principal components are plotted using the ade4 package in R with each sample represented by a filled circle. The centre of gravity for each cluster is marked by a rectangle and the coloured ellipse covers 67% of the samples belonging to the cluster. IBD, inflammatory bowel disease. **d**, Abundances of the main contributors of each enterotype from the Sanger metagenomes. See Fig. 1 for definition of box plot. **e**, Co-occurrence networks of the three enterotypes from the Sanger metagenomes. Unclassified genera under a higher rank are marked by asterisks in **b** and **e**.

to survive and persist in the human gut. This example illustrates that abundant species or genera cannot reveal the entire functional complexity of the gut microbiota. More reference genomes will facilitate better taxonomic assignment from samples and thus the detection of more low-abundance species. However, there is not much room for as yet undetected, abundant genera. Even with our limited genus assignment rate of 52.8% of all reads, we estimate that we miss another 30.7% of the already classified genera owing to our strict assignment criteria (Supplementary Fig. 1); that is, only 16.5% of all reads are likely to belong to hitherto unknown genera.

Detection of enterotypes, cross-national clusters

To get an overview of species variation we used phylogenetic profile similarities obtained by mapping metagenomic reads to the 1,511 reference genomes (Fig. 2a; see Supplementary Methods section 4.1). We excluded the two American Sanger-sequenced samples[4] from further analysis because of an unusual, very low fraction of Bacteroidetes and suspected technical artefacts[19]. Multidimensional cluster analysis and principal component analysis (PCA) revealed that the remaining 33 samples formed three distinct clusters that we designate as enterotypes (see Supplementary Notes section 4.1, Supplementary Fig. 3a and Supplementary Table 8). Each of these three enterotypes are identifiable by the variation in the levels of one of three genera: *Bacteroides* (enterotype 1), *Prevotella* (enterotype 2) and *Ruminococcus* (enterotype 3) (Fig. 2a, d), which was reproduced using independent array-based HITChip[20] data in a subset of 22 European samples (Supplementary Fig. 4 and Supplementary Notes section 4.5). The same analysis on two larger published gut microbiome data sets of different origins (16S pyrosequencing data from 154 American individuals[5] and Illumina-based metagenomics data from 85 Danish individuals[8]; Supplementary Methods section 5) shows that these data sets could also be represented best by three clusters (Supplementary Fig. 3b, c and Supplementary Tables 9, 10). Two of these are also driven by *Bacteroides* and *Prevotella*, whereas the third cluster is mostly driven by related groups of the order Clostridiales, *Blautia* and unclassified Lachnospiraceae in the 16S rDNA and Illumina data, respectively (Fig. 2b, c). This can be explained by a different reference data set in the instance of the 16S rDNA data, different mapping behaviour of short reads in the case of the Illumina data or current taxonomic uncertainties in the Lachnospiraceae and Ruminococcaceae clades (see Supplementary Notes section 4.2). The differences might also hint at community subpopulations within this enterotype, which might only be detectable with substantially more samples. Correlation analysis of the Sanger data revealed that abundances of each of the three discriminating genera strongly correlate (that is, they co-occur or avoid each other) with those of other genera (Fig. 2d; see Supplementary Methods section 11), indicating that the enterotypes are in fact driven by groups of species that together contribute to the preferred community compositions.

We demonstrate further the robustness of the enterotypes using two distinct statistical concepts. First, we used the silhouette coefficient[21] to validate that the three clusters are superior to clusterings obtained from various randomizations of the genus profile data, indicating a potential role for the interactions between co-occurring genera (see Supplementary Fig. 5 and Supplementary Notes section 4.3). Second, we used supervised learning and cross-validation to establish that these clusters have non-random characteristics that can be modelled and subsequently used to classify new samples (learning on clusters from randomized genus profiles led to considerably worse classification performance; see Supplementary Fig. 6 and Supplementary Notes section 4.4). These consistent results indicate that enterotypes will be identifiable in human gut metagenomes also from larger cohorts.

We then clustered the 33 samples using a purely functional metric: the abundance of the assigned orthologous groups (Fig. 3a). Remarkably, this clustering also showed a similar grouping of the samples with only minor differences (five samples placed in different clusters compared

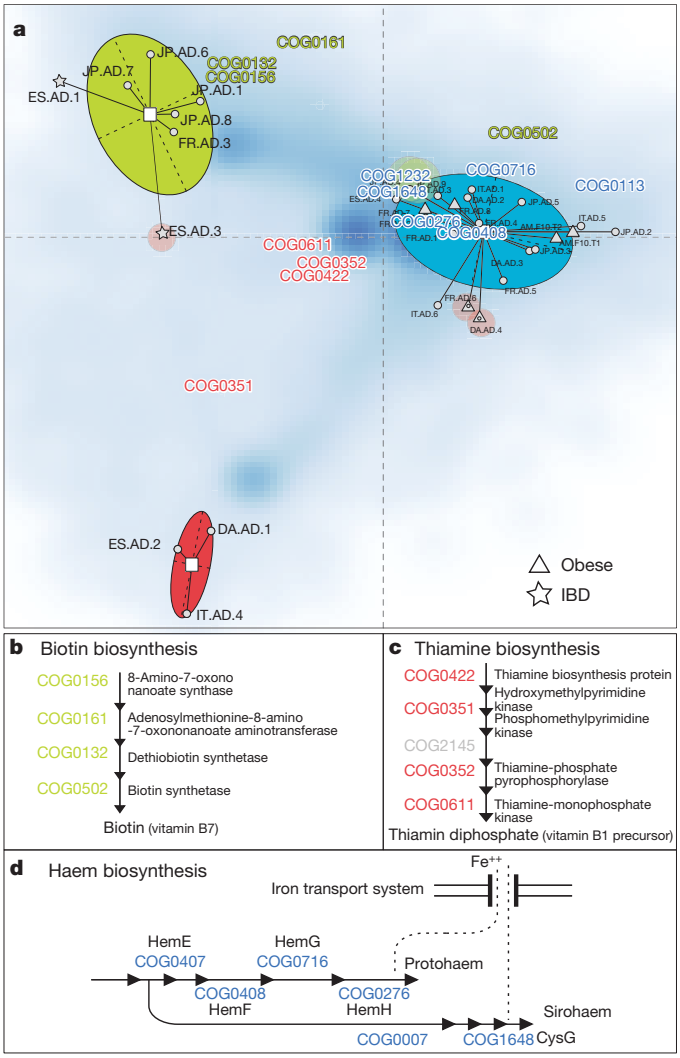

Figure 3 | Functional differences between enterotypes. **a**, Between-class analysis (see Fig. 2) of orthologous group abundances showing only minor disagreements with enterotypes (unfilled circles indicate the differing samples). The blue cloud represents the local density estimated from the coordinates of orthologous groups; positions of selected orthologous groups are highlighted. **b**, Four enzymes in the biotin biosynthesis pathway (COG0132, COG0156, COG0161 and COG0502) are overrepresented in enterotype 1. **c**, Four enzymes in the thiamine biosynthesis pathway (COG0422, COG0351, COG0352 and COG0611) are overrepresented in enterotype 2. **d**, Six enzymes in the haem biosynthesis pathway (COG0007, COG0276, COG407, COG0408, COG0716 and COG1648) are overrepresented in enterotype 3.

to Fig. 2a), indicating that function and species composition roughly coincide with some exceptions such as Spanish sample ES-AD-3, whose genus composition belongs to enterotype 2 whereas its functional composition is similar to members of enterotype 1. This individual has high levels of phage-related genes compared to the other samples (see Supplementary Fig. 7), hinting at partial temporal variability and dynamics of the microbiota, and perhaps indicating phage or virus bursts.

The robustness and predictability of the enterotypes in different cohorts and at multiple phylogenetic and functional levels indicates that they are the result of well-balanced, defined microbial community compositions of which only a limited number exist across individuals. These enterotypes are not as sharply delimited as, for example, human blood groups; they are, in contrast, densely populated areas in a multidimensional space of community composition. They are nevertheless likely to characterize individuals, in line with previous reports that gut

microbiota are quite stable in individuals and can even be restored after perturbation[22–25].

Variation between enterotypes

To determine the phylogenetic and functional basis of the enterotypes, we investigated in detail their differences in composition at the phylum, genus, gene and pathway level as well as correlations in abundance of co-occurring genera (Figs 2, 3; also see Supplementary Methods sections 10, 11 and 12). Enterotype 1, containing eight samples, is enriched in *Bacteroides* ($P < 0.01$; Supplementary Fig. 8), which co-occurs, for example, with *Parabacteroides* (see Supplementary Table 11 for enriched genera and Fig. 2e for correlation networks of co-occurring genera in each enterotype). The drivers of this enterotype seem to derive energy primarily from carbohydrates and proteins through fermentation, as these closely related genera have a very broad saccharolytic potential[26] and because genes encoding enzymes involved in the degradation of these substrates (galactosidases, hexosaminidases, proteases) along with glycolysis and pentose phosphate pathways are enriched in this enterotype (see Supplementary Tables 12, 13). Enterotype 2 contains six samples and is enriched in *Prevotella* ($P < 0.01$; Supplementary Fig. 9) and the co-occurring *Desulfovibrio*, which can act in synergy to degrade mucin glycoproteins present in the mucosal layer of the gut: *Prevotella* is a known mucin-degrader and *Desulfovibrio* may enhance the rate-limiting mucin desulphation step by removing the sulphate[27]. Enterotype 3 is the most frequent and is enriched in *Ruminococcus* ($P < 0.01$; Supplementary Fig. 10) as well as co-occurring *Akkermansia*, both known to comprise species able to degrade mucins[28]. It is also enriched in membrane transporters, mostly of sugars, indicating the efficient binding of mucin and its subsequent hydrolysis as well as uptake of the resulting simple sugars by these genera. The enriched genera indicate that enterotypes use different routes to generate energy from fermentable substrates available in the colon, reminiscent of a potential specialization in ecological niches or guilds. In addition to the conversion of complex carbohydrates into absorbable substrates, the gut microbiota is also beneficial to the human host by producing vitamins. Although all the vitamin metabolism pathways are represented in all samples, enterotypes 1 and 2 were enriched in biosynthesis of different vitamins: biotin (Fig. 3b), riboflavin, pantothenate and ascorbate in the former, and thiamine (Fig. 3c) and folate in the latter. These phylogenetic and functional differences among enterotypes thus reflect different combinations of microbial trophic chains with a probable impact on synergistic interrelations with the human hosts.

Functional biomarkers for host properties

Enterotypes do not seem to differ in functional richness (Supplementary Fig. 11), and virtually none of several measured host properties, namely nationality, gender, age or body mass index (BMI), significantly correlates with the enterotypes (with the exception of enterotype 1, which is enriched in Japanese individuals). However, some strong correlations do occur between host properties and particular functions, at the genes or module level (a module is a part of a pathway that is functionally tightly interconnected; see Supplementary Methods sections 6, 13 and Supplementary Notes section 6). The only significant correlation between a host property and a taxonomic group is a negative one between age and the abundance of an unknown Clostridiales genus ($P < 0.02$) containing three obligate anaerobes (Supplementary Fig. 12a; see Supplementary Notes section 6.2). It should be noted that age is not constant across the nationalities (in our data set, Italians are relatively old and Japanese young), but that individuals did not stratify by nationality, indicating that this is not a confounding factor. Our data did not reveal any correlation between BMI and the Firmicutes/Bacteroidetes ratio and we thus cannot contribute to the ongoing debate on the relationship between this ratio and obesity[29,30].

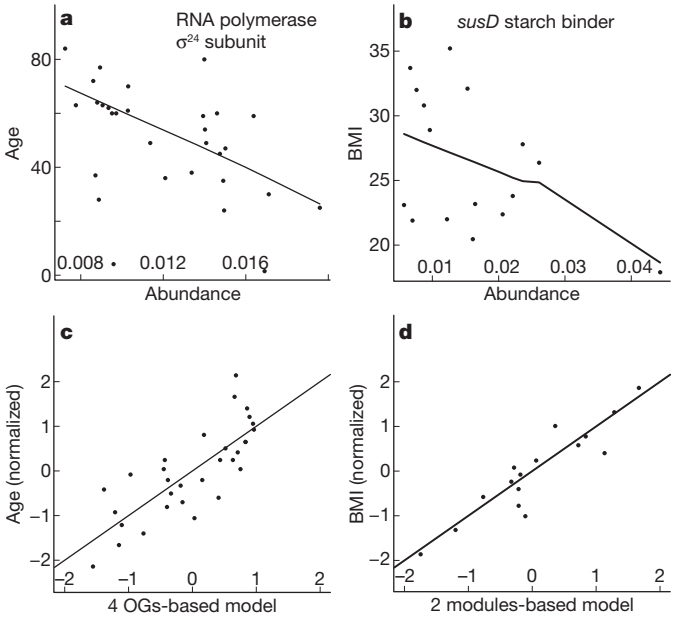

Figure 4 | Correlations with host properties. **a**, Pairwise correlation of RNA polymerase facultative σ^{24} subunit (COG1595) with age ($P = 0.03$, rho $= -0.59$). **b**, Pairwise correlation of SusD, a family of proteins that bind glycan molecules before they are transported into the cell, and BMI ($P = 0.27$, rho $= -0.29$, weak correlation). **c**, Multiple orthologous groups (OGs) (COG0085, COG0086, COG0438 and COG0739; see Supplementary Table 18) significantly correlating with age when combined into a linear model (see Supplementary Methods section 13 and ref. 40 for details; $P = 2.75 \times 10^{-5}$, adjusted $R^2 = 0.57$). **d**, Two modules, ATPase complex and ectoine biosynthesis (M00051), significantly correlating with BMI when combined into a linear model ($P = 6.786 \times 10^{-6}$, adjusted $R^2 = 0.82$).

In contrast to the minor phylogenetic signal, we found several significant functional correlations with each of the host properties studied (after correcting for multiple testing to avoid artefacts; see Supplementary Methods section 13), indicating that metagenomics-derived functional biomarkers might be more robust than phylogenetic ones. For example, the abundance of ten orthologous groups varies more between than within nationalities (Supplementary Table 14), although overall, the functional composition in total was remarkably similar among the nations (also with respect to the functional core; see Supplementary Fig. 13). For gender, we find five functional modules and one orthologous group that significantly correlate ($P < 0.05$; for example, enriched aspartate biosynthesis modules in males; see Supplementary Table 16). In addition, twelve orthologous groups significantly correlate with age (Supplementary Table 17). For instance, starch degradation enzymes such as glycosidases and glucan phosphorylases increase with age (which could be a reaction to decreased efficiency of host breakdown of dietary carbohydrates with age[31]) and so does the secA preprotein translocase (Supplementary Fig. 14). Conversely, an orthologous group coding for the facultative σ^{24} subunit of RNA polymerase, which drives expression under various stress responses and is linked to intestinal survival[32], decreases with age (Fig. 4a). One explanation for this could be the reduced need for stress response in the gut due to the age-associated decline in host immune response[33] (immunosenescence). Our analyses also identified three marker modules that correlate strongly with the hosts' BMI (Supplementary Table 19 and Supplementary Fig. 14), two of which are ATPase complexes, supporting the link found between the gut microbiota's capacity for energy harvest and obesity in the host[34]. Interestingly, functional markers found by a data-driven approach (derived from the metagenomes without previous knowledge) gave much stronger correlations than genes for which a link would be expected (for example, *susC*/*susD*, involved in starch utilization[26];

Fig. 4b). Linear models combining the abundance of only a few functional modules correlate even better with host properties (Fig. 4c, d). It should be noted that given the possibility of many confounding variables owing to the heterogeneity and size of our cohort, these observations will need to be substantiated using larger, independent cohorts in the future. Furthermore, patterns in metagenomics data can (partly) reflect indirect factors[9] such as genome size[35] (the smaller the average genome size of a sample, the higher the relative fraction of single copy genes therein), which, however, does not matter for diagnostics.

Although individual host properties do not explain the enterotypes, the latter might be driven by a complex mixture of functional properties, by host immune modulation or by hitherto unexplored physiological conditions such as transit time or pH of luminal contents. Furthermore, the three major enterotypes could be triggered by the three distinct pathways for hydrogen disposal[36] (Supplementary Notes section 6.4). Indeed, despite their low abundance, *Methanobrevibacter* (a methanogen) and *Desulfovibrio* (a known sulphate-reducer) are enriched in enterotypes 3 and 1, respectively.

Taken together, we have demonstrated the existence of enterotypes in the human gut microbiome and have identified three of them that vary in species and functional composition using data that spans several nations and continents. As our current data do not reveal which environmental or even genetic factors are causing the clustering, and as faecal samples are not representative of the entire intestine, we anticipate that the enterotypes introduced here will be refined with deeper and broader analysis of individuals' microbiomes. Presumably, enterotypes are not limited to humans but also occur in animals. Their future investigation might well reveal novel facets of human and animal symbiotic biology and lead to the discovery of those microbial properties correlated with the health status of individuals. We anticipate that they might allow classification of human groups that respond differently to diet or drug intake. Enterotypes appear complex, are probably not driven by nutritional habits and cannot simply be explained by host properties such as age or BMI, although there are functional markers such as genes or modules that correlate remarkably well with individual features. The latter might be utilizable for diagnostic and perhaps even prognostic tools for numerous human disorders, for instance colorectal cancer and obesity-linked co-morbidities such as metabolic syndrome, diabetes and cardiovascular pathologies.

METHODS SUMMARY

Sample collection. Human faecal samples from European individuals were collected and frozen immediately, and DNA was purified as described previously[37]. Sequencing was carried out by Sanger-sequencing random shotgun DNA libraries of 3 kb using standard protocols established at Genoscope. For sequence processing, cloning vector, sequencing primers and low-quality bases were end-trimmed from raw Sanger reads, and possible human DNA sequences were removed. Reads were processed by the SMASH comparative metagenomics pipeline[38] for assembly and gene prediction.

Informed consent was obtained from the 22 European subjects. Sample collection and experiments were approved by the following ethics committees: MetaHIT (Danish), ethical committee of the Capital Region of Denmark; MetaHIT (Spanish), CEIC, Hospital Vall d'Hebron; MicroObes, Ethical Committee for Studies with Human Subjects of Cochin Hospital in Paris, France; MicroAge, Joint Ethical Committee of the University of Camerino.

Phylogenetic annotation. Phylogenetic annotation of samples was performed by (1) aligning reads (Sanger/Illumina) against a database of 1,511 reference genomes (listed in Supplementary Table 3); or (2) classifying 16S rDNA reads using RDP classifier[39]. Genus and phylum abundance was estimated after normalizing for genome size for the former, and for 16S gene copy number for the latter.

Functional annotation. Genes were functionally annotated using BLASTP against eggNOG (v2) and KEGG (v50) databases. Protein abundances were estimated after normalizing for protein length. Functional abundance profiles at eggNOG, KEGG orthologous group, functional module and pathway level were created.

Clustering and classification. Samples were clustered using Jensen–Shannon distance and partitioning around medoid (PAM) clustering. Optimal number of clusters was estimated using the Calinski–Harabasz (CH) index. We used the silhouette validation technique for assessing the robustness of clusters. Additionally, within a cross-validation scheme, we trained predictive decision tree models on clusters obtained using the same clustering method and evaluated the classification of hold-out samples by accuracy, average precision and average precision gain.

Statistics. Correlations between metadata and feature abundances were computed as described previously[40], based on multiple-testing corrected pairwise Spearman correlation analysis and stepwise regression for multi-feature model building. For categorical metadata and enterotype comparisons, samples were pooled into bins (male/female, obese/lean, one enterotype/rest, specific nationality/rest etc) and significant features were identified using Fisher's exact test with multiple testing correction of *P* values.

Received 12 March 2010; accepted 18 December 2010.
Published online 20 April 2011.

1. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308,** 1635–1638 (2005).
2. Hayashi, H., Sakamoto, M. & Benno, Y. Phylogenetic analysis of the human gut microbiota using 16S rDNA clone libraries and strictly anaerobic culture-based methods. *Microbiol. Immunol.* **46,** 535–548 (2002).
3. Lay, C. *et al.* Colonic microbiota signatures across five northern European countries. *Appl. Environ. Microbiol.* **71,** 4153–4155 (2005).
4. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312,** 1355–1359 (2006).
5. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).
6. Kurokawa, K. *et al.* Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* **14,** 169–181 (2007).
7. Zoetendal, E. G., Rajilic-Stojanovic, M. & de Vos, W. M. High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. *Gut* **57,** 1605–1615 (2008).
8. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
9. Raes, J. & Bork, P. Molecular eco-systems biology: towards an understanding of community function. *Nature Rev. Microbiol.* **6,** 693–699 (2008).
10. Nelson, K. E. *et al.* A catalog of reference genomes from the human microbiome. *Science* **328,** 994–999 (2010).
11. MetaHIT Consortium. *MetaHIT Draft Bacterial Genomes at the Sanger Institute.* ⟨http://www.sanger.ac.uk/resources/downloads/bacteria/metahit/⟩ (9 July 2010).
12. Muller, J. *et al.* eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.* **38,** D190–D195 (2010).
13. Palmer, C., Bik, E. M., Digiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol.* **5,** e177 (2007).
14. Tap, J. *et al.* Towards the human intestinal microbiota phylogenetic core. *Environ. Microbiol.* **11,** 2574–2584 (2009).
15. Jensen, L. J. *et al.* STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **37,** D412–D416 (2009).
16. Dethlefsen, L., Huse, S., Sogin, M. L. & Relman, D. A. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol.* **6,** e280 (2008).
17. Walker, A. Say hello to our little friends. *Nature Rev. Microbiol.* **5,** 572–573 (2007).
18. Krogfelt, K. A. Bacterial adhesion: genetics, biogenesis, and role in pathogenesis of fimbrial adhesins of *Escherichia coli*. *Rev. Infect. Dis.* **13,** 721–735 (1991).
19. Salonen, A. *et al.* Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *J. Microbiol. Methods* **81,** 127–134 (2010).
20. Rajilic-Stojanovic, M. *et al.* Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environ. Microbiol.* **11,** 1736–1751 (2009).
21. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20,** 53–65 (1987).
22. Vanhoutte, T., Huys, G., Brandt, E., d. & Swings, J. Temporal stability analysis of the microbiota in human feces by denaturing gradient gel electrophoresis using universal and group-specific 16S rRNA gene primers. *FEMS Microbiol. Ecol.* **48,** 437–446 (2004).
23. Tannock, G. W. *et al.* Analysis of the fecal microflora of human subjects consuming a probiotic product containing *Lactobacillus rhamnosus* DR20. *Appl. Environ. Microbiol.* **66,** 2578–2588 (2000).
24. Seksik, P. *et al.* Alterations of the dominant faecal bacterial groups in patients with Crohn's disease of the colon. *Gut* **52,** 237–242 (2003).
25. Costello, E. K. *et al.* Bacterial community variation in human body habitats across space and time. *Science* **326,** 1694–1697 (2009).
26. Martens, E. C., Koropatkin, N. M., Smith, T. J. & Gordon, J. I. Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm. *J. Biol. Chem.* **284,** 24673–24677 (2009).
27. Wright, D. P., Rosendale, D. I. & Roberton, A. M. *Prevotella* enzymes involved in mucin oligosaccharide degradation and evidence for a small operon of genes expressed during growth on mucin. *FEMS Microbiol. Lett.* **190,** 73–79 (2000).
28. Derrien, M., Vaughan, E. E., Plugge, C. M. & de Vos, W. M. *Akkermansia muciniphila* gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. *Int. J. Syst. Evol. Microbiol.* **54,** 1469–1476 (2004).
29. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444,** 1022–1023 (2006).
30. Schwiertz, A. *et al.* Microbiota and SCFA in lean and overweight healthy subjects. *Obesity* **18,** 190–195 (2009).

31. Woodmansey, E. J. Intestinal bacteria and ageing. *J. Appl. Microbiol.* **102,** 1178–1186 (2007).
32. Kovacikova, G. & Skorupski, K. The alternative sigma factor σ^E plays an important role in intestinal survival and virulence in *Vibrio cholerae*. *Infect. Immun.* **70,** 5355–5362 (2002).
33. Fujihashi, K. & Kiyono, H. Mucosal immunosenescence: new developments and vaccines to control infectious diseases. *Trends Immunol.* **30,** 334–343 (2009).
34. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–1031 (2006).
35. Raes, J., Korbel, J. O., Lercher, M. J., von Mering, C. & Bork, P. Prediction of effective genome size in metagenomic samples. *Genome Biol.* **8,** R10 (2007).
36. Gibson, G. R. *et al.* Alternative pathways for hydrogen disposal during fermentation in the human colon. *Gut* **31,** 679–683 (1990).
37. Godon, J. J., Zumstein, E., Dabert, P., Habouzit, F. & Moletta, R. Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis. *Appl. Environ. Microbiol.* **63,** 2802–2813 (1997).
38. Arumugam, M., Harrington, E. D., Foerstner, K. U., Raes, J. & Bork, P. Smash Community: a metagenomic annotation and analysis tool. *Bioinformatics* **26,** 2977–2978 (2010).
39. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73,** 5261–5267 (2007).
40. Gianoulis, T. A. *et al.* Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc. Natl Acad. Sci. USA* **106,** 1374–1379 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors are grateful to C. Creevey, G. Falony and members of the Bork group at EMBL for discussions and assistance. We thank the EMBL IT core facility and Y. Yuan for managing the high-performance computing resources. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013): MetaHIT, grant agreement HEALTH-F4-2007-201052, EMBL, the Lundbeck Foundation Centre for Applied Medical Genomics in Personalized Disease Prediction, Prevention and Care (LuCAMP), Novo Nordisk Foundation and the International Science and Technology Cooperation Project in China (0806). Obese/non-obese volunteers for the MicroObes study were recruited from the SU.VI.MAX cohort study coordinated by P. Galan and S. Hercberg, and metagenome sequencing was funded by Agence Nationale de la Recherche (ANR); volunteers for MicroAge study were recruited from the CROWNALIFE cohort study coordinated by S. Silvi and A. Cresci, and metagenome sequencing was funded by GenoScope. Ciberehd is funded by the Instituto de Salud Carlos III (Spain). J.R. is supported by the Institute for the encouragement of Scientific Research and Innovation of Brussels (ISRIB) and the Odysseus programme of the Fund for Scientific Research Flanders (FWO). We are thankful to the Human Microbiome Project for generating the reference genomes from human gut microbes and the International Human Microbiome Consortium for discussions and exchange of data.

Author Contributions All authors are members of the Metagenomics of the Human Intestinal Tract (MetaHIT) Consortium. Jun W., F.G., O.P., W.M.d.V., S.B., J.D., Jean W., S.D.E. and P.B. managed the project. N.B., F.C., T.H., C.M. and T. N. performed clinical analyses. M.L. and F.L. performed DNA extraction. E.P., D.L.P., T.B., J.P. and E.U. performed DNA sequencing. M.A., J.R., S.D.E. and P.B. designed the analyses. M.A., J.R., T.Y., D.R.M., G.R.F., J.T., J.-M.B., M.B., L.F., L.G., M.K., H.B.N., N.P., J.Q., T.S.-P., S.T., D.T., E.G.Z., S.D.E. and P.B. performed the analyses. M.A., J.R., P.B. and S.D.E. wrote the manuscript. M.H., T.H., K.K. and the MetaHIT Consortium members contributed to the design and execution of the study.

Author Information Raw Sanger read data from the European faecal metagenomes have been deposited in the NCBI Trace Archive with the following project identifiers: MH6 (33049), MH13 (33053), MH12 (33055), MH30 (33057), CD1 (33059), CD2 (33061), UC4 (33113), UC6 (33063), NO1 (33305), NO3 (33307), NO4 (33309), NO8 (33311), OB2 (33313), OB1 (38231), OB6 (38233), OB8 (45929), A (63073), B (63075), C (63077), D (63079), E (63081), G (63083). Contigs, genes and annotations are available to download from http://www.bork.embl.de/Docu/Arumugam_et_al_2011/. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.B. (bork@embl.de) or S.D.E. (dusko.ehrlich@jouy.inra.fr).

MetaHIT Consortium (additional members)

María Antolín[1], François Artiguenave[2], Hervé M. Blottiere[3], Mathieu Almeida[3], Christian Brechot[12], Carlos Cara[4], Christian Chervaux[5], Antonella Cultrone[3], Christine Delorme[3], Gérard Denariaz[5], Rozenn Dervyn[3], Konrad U. Foerstner[6,7], Carsten Friss[8], Maarten van de Guchte[3], Eric Guedon[3], Florence Haimet[3], Wolfgang Huber[6], Johan van Hylckama-Vlieg[5], Alexandre Jamet[3], Catherine Juste[3], Ghalia Kaci[3], Jan Knol[5], Omar Lakhdari[3], Severine Layec[3], Karine Le Roux[3], Emmanuelle Maguin[3], Alexandre Mérieux[12], Raquel Melo Minardi[2], Christine M'rini[12], Jean Muller[9], Raish Oozeer[5], Julian Parkhill[10], Pierre Renault[3], Maria Rescigno[11], Nicolas Sanchez[3], Shinichi Sunagawa[6], Antonio Torrejon[1], Keith Turner[10], Gaetana Vandemeulebrouck[3], Encarna Varela[1], Yohanan Winogradsky[3] & Georg Zeller[6]

[1]Digestive System Research Unit, University Hospital Vall d'Hebron, Ciberehd, 08035 Barcelona, Spain. [2]Commissariat à l'Energie Atomique, Genoscope, 91000 Evry, France. [3]Institut National de la Recherche Agronomique, 78350 Jouy en Josas, France. [4]UCB Pharma SA, 28046 Madrid, Spain. [5]Danone Research, 91120 Palaiseau, France. [6]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [7]Heidelberger Strasse 24, 64285 Darmstadt, Germany. [8]Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark. [9]Institute of Genetics and Molecular and Cellular Biology, CNRS, INSERM, University of Strasbourg, 67404 Illkrich, France. [10]The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [11]Istituto Europeo di Oncologia, 20100 Milan, Italy. [12]Institut Mérieux, 17 rue Burgelat, 69002 Lyon, France.