

Annotation of the Human Genome by High-Throughput Sequence Analysis of Naturally Occurring Proteins

Simon J. McGowan[#], Jonathan Terrett[#], Clive G. Brown[#], Paul J. Adam, Louise Aldridge, Jason C. Allen, Bob Amess, Kristian A. Andrews, Martin Barnes¹, David E. Barnwell, Joanne Berry, Helen Bird, Robert S. Boyd, Marissa J. Broughton, Alice Brown, Jim A. Bruce, Luc M. J. Brusten, Nicholas J. Draper¹, Beverley M. Elsmore, Colin D. Freeman, David M. Giles, Haiping Gong, Darren Gormley, Matthew R. Griffiths, Tim D.R. Hawkes, Paul S. Haynes, Kate J. Heesom, Athula Herath, Katherine Hollis, Lindsey J. Hudson, Janet Inman, Merrill Jacobs, Darren Jarman, Imran Kibria, John J. Kilgour, Samuel K. Kinuthia, Kim E. Lane, Margaret L. Lees, Julie Loader, Andrew Longmore¹, Michael McEwan, Alice Middleton, Stephen Moore, Carol Murray, Helen M. Murray, C. Paul Myatt, Stanley S. Ng, Andrew O'Neil, Raj B. Parekh, Ashok Patel, Kaajal B. Patel, Sonal Patel, Thakor P. Patel, Robin J. Philp, Albert E. Platt, Helen Poyser, Cynthia Prendergast, Sally Prime, Nicholas Redpath, Mike Reeves, Andrew W. Robinson, Christian Rohlf, Jeffrey M. Rosenbaum, Martin Schenker, Elaine Scrivener, Nigel Shipston, Shaistah Siddiq, Christopher Southan, Daniel I. R. Spencer, Alasdair Stamps, Marc A. Steffens, David Stevenson, Gavin M.A. Sweetman, Stephen Taylor, Reid Townsend, Andrew M. Ventom, Martin N. H. Waller¹, Celia Weresch, Amanda M. Williams, Richard J. Woolliscroft, Xiaohong Yu and Andrew Lyall*

Oxford GlycoSciences plc, The Forum, 86 Milton Park, Abingdon, OX14 4RY, UK; ¹Tessella Support Services plc, 3 Vineyard Chambers, Abingdon, OX14 3PX, UK

Abstract: The identification of protein-coding genes is currently based on the merging of evidence and predictions from a variety of databases that may themselves contain inaccurate and partial information. We have developed a method for mapping accurate interpretations of protein MS-MS data to the genome. This approach enables verification of genes, exons, transcripts and variant transcripts as well as the *de novo* discovery of novel protein-coding genes. Here we describe improvements in spectral interpretation algorithms, multiple separation techniques, sub-cellular fractionation and novel bioinformatics approaches to characterise more than 14,000 naturally occurring human genes.

Key Words: Proteomics, Genome, Human, Genes, Protein Isoforms, Bioinformatics.

INTRODUCTION

Recent dramatic advances in defining the nucleotide sequence of the human genome have led to the near completion of this immense task (Venter *et al.*, 2001; International Human Genome Sequencing Consortium, 2001). There is little doubt that this sequence information will have a substantial impact on our understanding of many biological processes, including molecular evolution, comparative genomics, pathogenic mechanisms and molecular medicine. For the full medical value inherent in the sequence of the human genome to be realised, the 'organised' genome needs to be annotated (International Human Genome Sequencing Consortium, 2001; Ensembl). By this, is meant at least the following three things (i) the unambiguous identification of those regions of each chromosome that contain genes (ii) description of the exon fine structure of these genes and (iii) determination of the various protein products derived from each gene under different environmental conditions / tissue backgrounds. A subset of these genes / proteins will be involved in the

molecular basis of most if not all pathologies. Therefore, an important and immediate goal for the pharmaceutical industry is to identify all such genes in the human genome such that those relevant to disease can be analysed further.

Given that a small fraction of the genome appears to contain protein-coding genes, one approach to achieving the goal of identifying such genes has been to rely on clustered ESTs (de Souza *et al.*, 2000). While it is accepted that ESTs have contributed substantially to gene discovery, the reliance on EST clusters to reveal genes for naturally occurring proteins is based on unqualified assumptions, which have been discussed elsewhere (Aparicio, 2001). In particular, it is extremely difficult using this technique to distinguish real protein coding genes from silent-genes and pseudo-genes. The fact that a cDNA derived from an EST (or an EST cluster) can be used to generate a recombinant protein in a cellular system, says nothing about the natural occurrence of that protein. It is only possible to prove the existence of a protein with techniques such as direct protein analysis and thus 'authenticate' the cDNA. Comprehensive attempts using DNA microarrays (Shoemaker *et al.*, 2001) have been made to demonstrate expression of mRNAs corresponding to gene predictions, yet this still does not confirm that proteins are produced from these 'genes'.

*Address correspondence to this author at the Oxford GlycoSciences plc, The Forum, 86 Milton Park, Abingdon, OX14 4RY, UK; Tel: +44 (0) 1235 208000; Fax: +44 (0) 1235 208011; E-mail: andrew.lyall@ic4life.net

[#]These authors contributed equally to this work.

The sequence alignment methods used currently for genome annotation have created an extremely well annotated genome in a very short time, but are approaching a plateau in efficiency. High throughput EST sequencing has covered most tissues but there are many predicted genes that have no EST coverage. Even as the gene prediction algorithms become more sophisticated, each of these predicted genes usually requires confirmation via RT-PCR in an ordered and labour intensive manner. Once again however, this approach does not confirm that the transcript produces a natural protein or that the predicted transcript is accurate.

Finally, the number of protein-coding genes predicted using ESTs is considerably greater than the current consensus (Liang *et al.*, 2000). Assuming that the current consensus number of between 35,000 and 60,000 protein coding genes, is close to the truth (Liang *et al.*, 2000; Ewing and Green, 2000) then even if every protein-coding gene is represented within the EST databases, many of the remaining ESTs would not be expected to derive from genes coding for naturally occurring proteins. For at least these reasons, the goal of identifying and organising the protein-coding genes in the genome will only be achieved in the short term by amino-acid sequence analysis of naturally expressed proteins. This has so far been technically impractical and to date less than 1,000 naturally expressed human proteins (and their genes) have been unambiguously identified by direct peptide sequence analysis and deposited in the SwissProt protein sequence database. Amino-terminal sequence analysis of peptides by Edman degradation is relatively slow and too insensitive to deal economically with this task. Mass spectrometric methods to obtain peptide sequence information have been limited by the lack of algorithms to interpret *ab initio* tandem mass spectra (of the quality and signal:noise usually obtained), or to identify from the tandem fragmentation spectrum the nucleotide sequence encoding the peptide in a search space the size of the human genome.

There are further advantages to proteomics based genome annotation. Firstly, proteomics can utilise body fluids and cell types such as neutrophils and erythrocytes that have little RNA turnover and where EST sequencing struggles to identify rare transcripts. Secondly, cells can be sub-fractionated prior to proteomic analysis allowing enrichment of rare proteins (e.g. certain transmembrane receptors). Furthermore, proteomic analyses are not subject to the C-terminal (3') bias that is symptomatic of EST libraries derived from oligo dT primed reverse transcription.

Recently the concept of mapping peptide microsequences to genomic sequence data has been demonstrated (Choudhary *et al.*, 2001; Kuster *et al.*, 2001) and although this highlighted the benefits of finding potentially novel proteins that are not present in other databases, problems in mapping peptides that are derived from more than one exon are also noted. In this paper we describe how proteomics technologies have now reached the sensitivity of genome and cDNA sequencing, allowing a new set of proteins to be annotated, conceptual translations to be confirmed, and the subset of predicted genes that actually encode real proteins to be defined.

MATERIALS AND METHODS

Naturally occurring proteins analysed in our work were obtained from a wide variety of sources including some that have been subjected to sub-cellular fractionation prior to analysis: clinical cancer material (including breast, colon, prostate, leukaemia), cancer cell lines (including prostate, breast, colon, liver, kidney, pancreas, lung, prostate) B cells, T cells, serum, CSF, synovial fluid, fibroblasts, adipocytes, clinical normal material (including CNS, liver, breast, endothelial cells). Trypsin was obtained from Promega (Promega UK Ltd., Southampton, United Kingdom). Peptide pools were analysed by MALDI-MS using a Voyager-DE-STR equipped with a 337 nm laser (Applied BioSystems, Framingham, MA) and by ESI-MS/MS using a Q-TOF instrument (Micromass, Manchester, United Kingdom) filled with a nano-electrospray Z source.

Purification of Naturally Occurring Proteins

Cellular protein extracts were obtained as previously described (Page *et al.*, 1999). Protein extracts and peptide pools were generated 2-D SDS-PAGE (Page *et al.*, 1999) or by 1-D SDS-PAGE (Adam *et al.*, 2003), or peptide pools were created by the ICAT analysis (Gygi *et al.*, 1999).

Mass Spectrometry

The peptide pools were divided for Matrix-assisted laser-desorption time-of-flight analysis (MALDI-TOF) and electrospray analysis (see below), 20% and 80%, respectively. For MALDI-TOF analysis, the samples were purified prior to spotting on the target using a C18 cartridge (ZipTip, Millipore, Watford, United Kingdom). The samples were applied to the cartridge and washed three times (10 μ l each) with an aqueous solution containing 2.5% (v/v) acetonitrile. The peptides were then eluted directly onto the MALDI target with a solution (0.6 μ l) containing premixed matrix (alpha-cyano-4-hydroxycinnamic acid; Hewlett Packard GmbH, Boeblingen, Germany). Spectra were acquired on a Voyager-DE-STR equipped with a 337 nm laser (Applied BioSystems, Framingham, MA). The acceleration voltage was 20 kV with an extraction delay time of 125 ns. Data from a total of 500 laser pulses in 5 different spot locations were acquired. The data were initially processed using Applied BioSystems DataExplorer software consisting of baseline subtraction, noise reduction (set at 2), de-isotoping, internal calibration using 5 trypsin peaks (minimum acceptance is 3 peaks within 10 ppm), and export of peak list. Masses corresponding to trypsin peptides (produced by auto-catalysis or fragmentation in the mass spectrometers) and contaminant peptides from the gel-proteolysis process were removed from the exported peak list.

The remainder of each peptide pool was analyzed using either nanospray MS or capillary reversed-phase LC-MS/MS. The customized chromatograph consisted of a FAMOS autosampler (LC Packings, Camberley, United Kingdom), Rheos 4000 gradient pumps (Presearch, Hitchin, United Kingdom), and a minaturised C18 column prepared in house (estimated bed volume, 100 μ l) fitted with a 5 micron i.d. fused silica emitter (New Objective, Cambridge, MA). The peptides were loaded onto the equilibrated column

(1%, v/v, formic acid) and eluted (200 nl/min) with a linear gradient of 60%, v/v, acetonitrile over 20 min. The chromatograph was interfaced to an orthogonal quadrupole time-of-flight mass spectrometer (Q-TOF) using an electrospray Z source (Micromass, UK). The needle voltage was set at positive 1600-1800 V. Fragmentation spectra were acquired at the selected m/z values, which were determined from the MALDI-TOF analysis, principally according to the criterion of signal intensity.

Sequence Databanks

Sequence and annotation data were obtained from the following sequence databanks: NR (Oct 2002; <http://www.ncbi.nlm.nih.gov/>), SwissProt (40.27; <http://us.expasy.org/sprot/>), IPI (2.12; <http://www.ebi.ac.uk/IPI/IPIhelp.html>), TREMBL (Aug 2002; <http://www.ebi.ac.uk/trembl/>). The public domain human genome was that of Ensembl build 30. The polymorphic 'hypothetical transcriptome' was constructed from transcripts predicted by FGENES, FGENESH (Softberry Inc, Mount Kisco, NY, USA), GENSCAN (Stanford University, Stanford, CA, USA) and genes and EST-genes from Ensembl build 30. We used SPLM (Softberry Inc, NY, USA) to create variant exon boundaries within this transcriptome. Unique SNP's were mapped using NCBI dbSNP.

Interpretation of Tandem Mass Spectra

We used the MTM (Robinson and Townsend, 2002), SEQUEST (Eng *et al.*, 1994) and Lutefisk (Taylor and Johnson, 1997) algorithms to interpret the high quality MS-MS tandem spectra generated in this work. We used a search space comprising NR, SwissProt, TREMBL and the polymorphic 'hypothetical transcriptome' sequences as the search space for SEQUEST. We used a database of *in silico*-derived tryptic peptides arising from these same sequences when using the MTM algorithm.

All currently available spectral interpretation software is prone to two kinds of error, those arising from inaccuracy and those arising from insensitivity. Inaccuracy causes incorrect interpretation of spectra. These are referred to as false positives. Insensitivity causes failure to interpret spectra that could be interpreted by a human expert. These are referred to as false negatives. Typically false positives are eliminated by manual inspection; a human expert categorizes the spectral interpretations as correct, incorrect and impossible to determine. The elimination of false negatives would require the manual inspection of large numbers of low quality spectra, something that is generally not regarded as a useful exercise. As these approaches are not logistically feasible for a high-throughput system such as we describe, here we have devised an automated system that reduces these problems to acceptable levels whilst requiring minimal human intervention. In addition we have provided users who are not expert in mass spectrometry with an interface that allows them to utilize their prior knowledge about biochemistry when considering our spectral interpretations. In the context of this work, these spectral interpretations, and the gene fine structures generated from them, are associated with gene predictions, cDNAs and knowledge about the separation techniques used to prepare

the samples. In our experience this has dramatically increased the utility of the human genome and we have exploited these methods in the discovery of drug targets and bio-markers in oncology (Adam *et al.*, 2003) and in CNS disorders Rohlff and Southan, 2002). The high throughput nature of our process mitigates the problem of insensitivity by generating extremely large numbers of high quality spectra and by ensuring that most peptides are submitted for mass spectrometry many times. We deal with false positives by using several orthogonal algorithms in such a manner as to complement each other. The first method (MTM) is of our own devising and is described elsewhere (Robinson and Townsend, 2002). The second method (SEQUEST) is in use in most proteomics laboratories in the world and its performance is well understood. We use it here with relatively stringent settings of the parameters (raw C > 1.1, delta Cn > 0.15) and a restricted search space (we only deploy it against the polymorphic, hypothetical transcriptome and genomic DNA around verified genes - not the full translation of the genomic DNA). The resulting peptide interpretations are subsequently evaluated using a count of the total number of y -ions matched, the length of the longest contiguous chain of y -ions and the cross-correlation between the original spectrum and a theoretical spectrum based on the peptide amino acid sequence. This is an automated process carried out by an algorithm developed at Oxford Glyco-Sciences (AutoDA). In addition, we have recently adopted Lutefisk and again use this against a restricted search space in the same manner. We used these methods to generate complete and complementary sets of spectral interpretations to feed into the gene building algorithm. Each set has been generated in a manner that minimizes its specific false positive rates and by combining them with the gene building algorithm, this is reduced further.

RESULTS AND DISCUSSION

In order to map microsequenced peptides derived from proteins to the human genome, we have combined new algorithms for the high-throughput precise interpretation of tandem MS spectra (Robinson and Townsend, 2002), with a dataset containing cDNAs, proteins, gene predictions and including SNPs, as well as all 6 reading frames of the human genome sequence. Although most of the peptide sequences (derived from approximately 212,000 MS/MS spectra) matched to the first dataset, more than 50 uniquely map to the 6 frame genomic translations.

The approach we have used (Fig. 1) comprises four components. (i) Integrating multiple protein separation platforms including cell fractionation, 1D and 2D gels, and ICAT (Gygi *et al.*, 1999) (ii) Directed tandem mass spectrometry (Adam *et al.*, 2003), and multiple algorithms to interpret the resulting fragmentation spectra. (iii) Using the peptide 'signatures' so obtained to search the publicly available human genome and other databases to identify transcripts and protein coding genomic loci. (iv) Applying bioinformatics to provide an increased peptide search space, map exon-crossing peptides, map peptides derived from polymorphic DNA that differs from the published human genome and to integrate all available information on the protein, and its corresponding features, in order to identify,

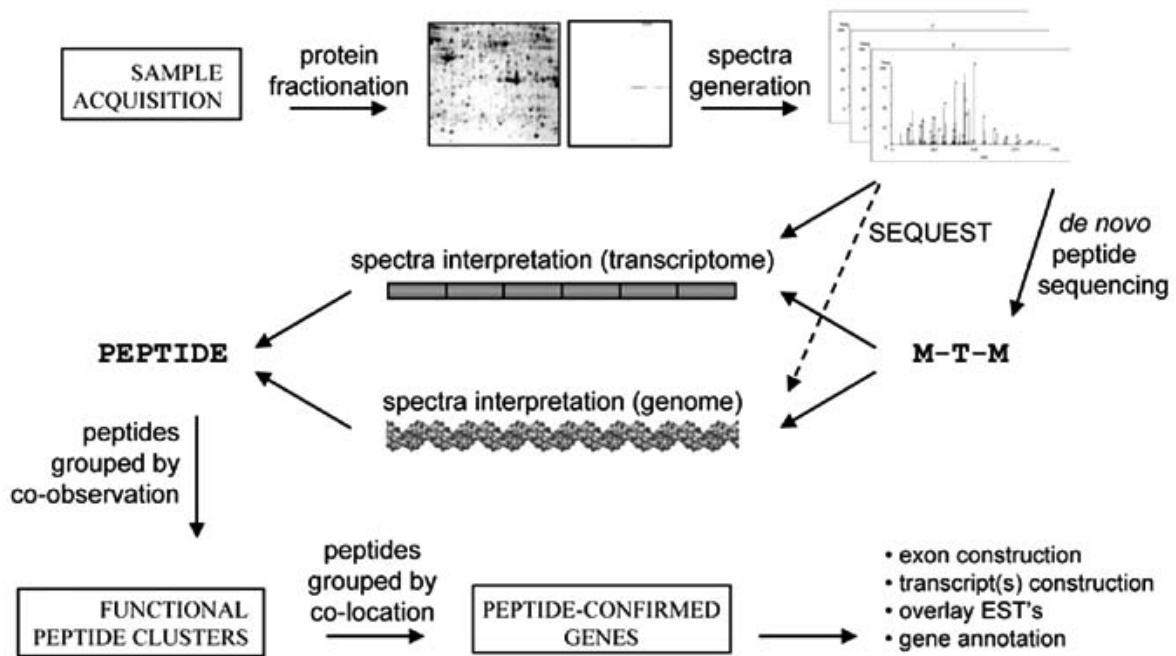


Fig. (1). Biological and bioinformatic processes used to construct the protein Atlas of the human genome via MTM and SEQUEST interpretation of MS-MS data. Fractionated and microsequenced proteins are applied to the transcriptome and the genome via MTM and SEQUEST interpretation of MS-MS-derived spectra. Identified proteins are used to interpret associated MALDI-TOF spectra. All genes/proteins are then clustered by co-observation of both sets of peptides (microsequenced and mass-observed) within the same fractionated proteins and by co-location within the genomic DNA.

locate and in many cases organise the gene coding for that protein into its constituent exons and transcripts.

In the context of the human genome and in searching for novel genes, the most comprehensive search space must be defined. However, it is not possible to simply apply the usual algorithms (e.g. SEQUEST, (Eng *et al.*, 1994)) to translations in all six reading frames of the human genome – the resulting dataset is too large to allow searching in a reasonable time and additionally its complexity will result in too many false interpretations of the spectrum. Instead, we have developed a novel method of *de novo* spectral interpretation to allow us to utilise this broadened search space. The *de novo* interpretation is carried out in a number of algorithmic stages. An initial algorithm is used to create a ‘signature’ for each spectrum. The signature consists of a read of a minimum of 3 consecutive amino acids and the N and C terminal masses that make up the complete peptide. This signature created is thus an MTM (mass (Nter) – trimer (sequence) – mass (Cter)) and is then searched against a database of conceptual tryptic peptides derived from the genome to produce a list of candidate explanations for the spectrum. This step dramatically reduces the search space for this spectral interpretation and subsequently allows a set of more exhaustive algorithms to eliminate all but one of the possible alternatives. The database of theoretical tryptic peptides is created to allow the linking of a given peptide sequence to one or more potential transcripts. Thus the elucidation of peptide sequences of spectra from a given peptide pool further defines the search space to the identified “proteins” for subsequent spectra derived from the same

protein(s). A final step therefore, allows the decoration of this identified database transcript using further spectra. All remaining MS-MS spectra for which a signature could not be unambiguously determined from the large conceptual set of potential peptides are now searched against only the peptides of the MS-MS identified proteins. This step removes the ambiguity in the peptide interpretation of such spectra – either they will have a single peptide interpretation derived from the identified proteins in the same peptide pool, or they will have no interpretation and will be discarded.

Mass spectrometry of proteins at Oxford GlycoSciences is a two stage process, in which MS-MS analysis is preceded by MALDI-TOF analysis of the same peptide pool (see Materials and Methods). This arrangement allows the identification and screening of contaminant peptides (for example, derived from trypsin) prior to analysis of the peptides using MS-MS and ensures that for all MS-MS spectra generated, there is also an associated MALDI-TOF spectrum. Therefore, once a database transcript has been identified by MS-MS spectral interpretation, peptide mass fingerprint of this transcript is used to verify the matching with masses identified in the related MALDI-TOF spectrum.

One of the keys to this high-throughput integration of genetics and proteomics has been the improvements in bioinformatics that we have implemented and in particular the development and use of the hypothetical transcriptome. This dataset contains a wide variety of known and hypothetical transcripts produced by gene prediction algorithms and by Ensembl (www.ensembl.org). Translation

and conceptual tryptic digestion of these sequences allows all potential MS-observable tryptic peptides to be added to the theoretical tryptic peptides derived from 6-frame translation of the human genome. Where a peptide is derived from more than one exon within a transcript (an exon-crossing peptide), both genomic locations that give rise to the peptide are recorded in the database. Correlation of the positions of known SNPs with the DNA that gave rise to these peptide sequences allows the generation of all possible variant peptides and again, these are added to the observable tryptic peptides database. This dataset is a first approximation of the optimal search space for spectra derived from human tissue and is an attempt to estimate all observable tryptic peptides that are present in the human proteome (excluding post-translational modifications (PTMs)). As such, it may be used for automatically mapping peptides derived from exon-crossing peptides, as well as from transcripts that differ from the published genomic sequence by one or more single nucleotide polymorphisms.

The term 'transcriptome' has already been used elsewhere (Caron *et al.*, 2001) to describe those genes that are expressed within a defined tissue type. In seeking to name this dataset, we have extended this definition to include all possible genes, transcripts and their variations that could be expressed in any tissue background and we therefore, refer to this dataset as a 'polymorphic, hypothetical transcriptome'. In mimicking the molecular biology within the cell, this dataset is a logical and complementary addition to the genomic DNA translation dataset.

The transcriptome dataset is over-predicted and will certainly contain predictions that are not protein-coding genes. It has however, minimised exposure of the spectral interpretation algorithms to non-coding, extragenic DNA (likely to be > 95% of the genome), and thereby reduced the level of false interpretation of spectra when attempting to use SEQUEST (or other correlative methods) in combination with a genome search space.

We have recently extended this methodology to allow SEQUEST to carry out limited searches of translations of non-exonic genomic-DNA. Where MTM or SEQUEST has identified a peptide from one of the predicted transcripts, we regard the gene prediction as verified within the transcriptome dataset. It is likely that any uninterpreted spectra from the same protein isolate, for example a spot from a 2D gel, are derived from peptides that originate from a translation of the verified gene (or from other similarly validated regions of the genome in the case of a mixed protein isolate). The algorithms may have failed to interpret the spectra because the peptide was not in the available search space, i.e. it was derived from a translation of an exon that was either not predicted or that was predicted with incorrect exon boundaries. To overcome this we allow the algorithms to attempt subsequent interpretation of these spectra with an expanded dataset based on the region of the genome underlying the validated transcript. This focused dataset is generated from all intronic DNA under the verified gene prediction as well as short regions both upstream and downstream of the prediction. In addition, exon boundaries may be varied algorithmically and exons may be paired together to generate new hypothetical exon-crossing peptides

(within the constraints of maintaining phase of the translated gene product). SNPs that are proximal to exon boundaries must also be taken into account because the polymorphism may result in the creation or destruction of exon boundaries. Splice variants are increasingly being recognised as an important and subtle modulator of protein functionality (Modrek and Lee, 2002) and this approach has allowed us to discover potential variants of the MS/MS-verified genes.

Using this analysis pipeline, we have mapped more than 212,000 MS/MS spectra to the human genome as well as approximately 950,000 MALDI mass-matched peptides from a wide range of tissues and disease states to annotate the human genome with 14,223 genes.

After peptide interpretations and genomic coordinates have been generated for all tandem spectra from a given peptide pool, they are algorithmically (and automatically) combined to yield a gene call for the protein(s). An initial pass is made to group all peptides that are found within the same members of the polymorphic transcriptome. Gene calls are then made independently for each of the subsequent groups. Masses arising from MALDI mass-spec analysis from the given peptide pool are mapped to members of the transcriptome on which the tandem peptide(s) from that pool have already been observed. The genes are constructed by linking co-observed peptides (peptides detected in the same pool) that are located on the same strand of the same chromosomal DNA and within a distance on the physical map, which is not beyond what would reasonably be considered as a maximum inter-exon distance ($\leq 80,000$ bp). Exon boundaries are fitted around these peptides using the 'pre-annotated' nature of the transcriptome. Where peptides have been mapped to the genomic DNA, but not to the transcriptome, exon boundaries are estimated around the peptides using a statistical algorithm based upon consensus splice sites and gene structures. Finally, all observations of each gene are consolidated to construct a likely transcript for that gene. This is attempted using the peptides, validated exons, surrounding and intervening exons and, if available, the inferred mass of the intact protein for each observation. Where required, potential splice variants are created to explain all possible MS-MS data for every gene.

Using the method described, 14,223 annotated genes have been arranged onto the human genome using Ensembl (build 30) as a framework. Figure (2) shows the process described in this paper in reverse and how drilling down beyond the chromosomal data highlights the coverage of protein coding genes (here illustrated using chromosome 19), the number of exons per gene, and finally how each tandem spectrum is aligned with genomic DNA sequence and the translated peptide sequence. The other chromosomes show a similar distribution of genes and similar coverage of the Ensembl estimates of gene counts. Ensembl (build 30) contained 22920 genes. On a chromosome-by-chromosome basis our protein sequencing has generated a consistent > 46% overlap with these loci, thereby eliminating the ambiguity for more than 10,500 predicted Ensembl genes. In addition however, we have protein sequence evidence for a further of about 3700 genes that were not annotated in Ensembl. To underline the value that comprehensive peptide sequencing brings to genome annotation we have further

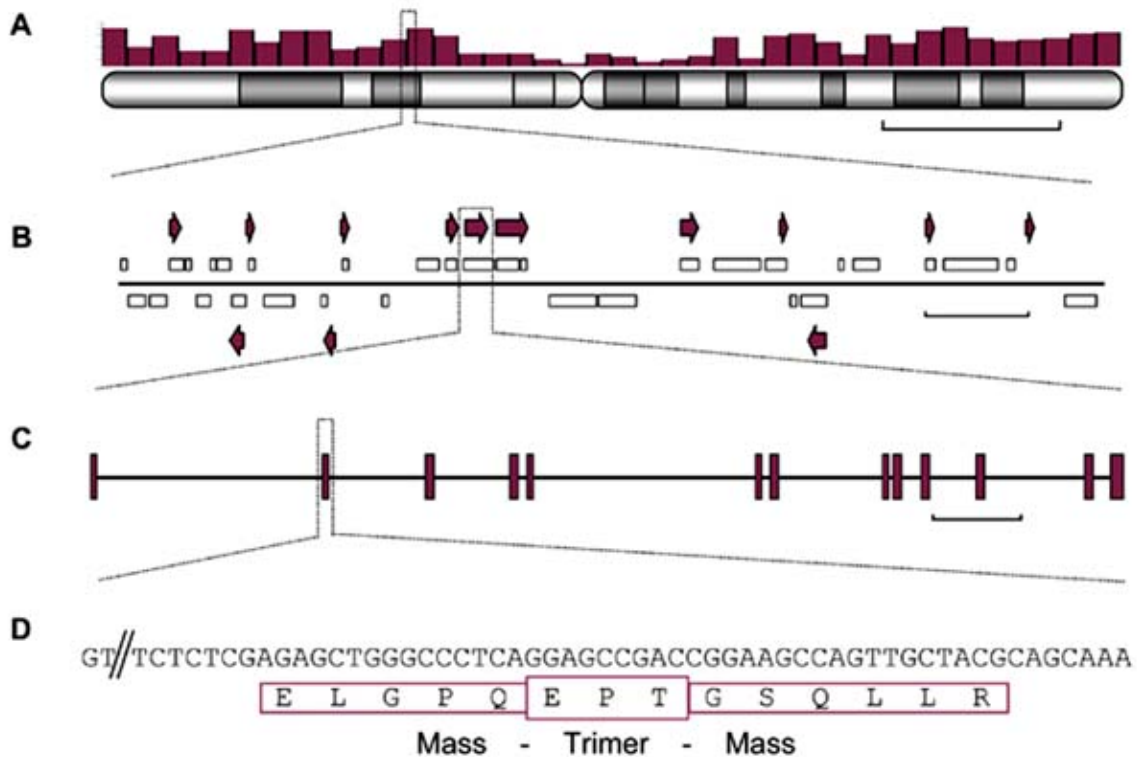


Fig. (2). Protein annotation of chromosome 19. A, Chart shows relative distribution of the 591 protein-coding genes identified in this study on chromosome 19 (line bar = 10,000,000 bp; height of vertical bars is proportional to number of genes). B, Protein-coding genes identified in this study (arrows) and Ensembl genes (rectangles) near the novel protein, C1orf24 (at 19p13.12) (line bar = 100,000 bp). C, Exon/intron structure of C1orf24 (line bar = 1,000 bp). D, Fine structure of C1orf24 exon 2 showing the MTM aligned directly to the genomic DNA sequence and protein translation.

characterised one such annotation representing those gene categories that are the most difficult to validate.

Validation of a Novel Predicted Protein

The process used to identify protein-coding genes and their exons is illustrated using a protein initially isolated from purified membrane preparations derived from chronic lymphoid leukaemia (CLL) samples and resolved using 1D gels (Fig. 3). Tryptic peptide pools were obtained from random and successive cuts from the entire gel. Microsequencing by tandem mass spectrometry of a peptide (ELGPQEPTGSQLLR) from a gel slice at about 80 kDa identified a GENSCAN predicted protein C1orf 24 from our over-predicted transcriptome. The predicted gene spanned 21 exons. Sequence database searching identified several public domain ESTs that showed identity to different regions of C1orf 24, but none of these ESTs overlapped with the sequenced peptide. These data did not indicate whether the GENSCAN prediction represented a single protein or multiple proteins.

Alignment of the predicted C1orf 24 protein with genomic DNA allowed the design of multiple PCR primers covering the putative C1orf 24 coding region. These primers were used to clone C1orf 24 from Daudi cell line cDNA. The largest PCR amplified clone was approximately 2.3 kb DNA sequence analysis of each amplified fragment revealed that

C1orf 24 represented multiple transcripts and encoded at least 4 proteins derived from alternative splicing and potentially alternative start codons. A total of 17 exons were identified from the alternatively spliced transcripts, allowing a contiguous conceptual protein to be created. Masses obtained from the same protein fragment as the microsequenced peptide were then positioned on this conceptual translation. Of 25 theoretically observable peptides (masses 750 – 2500), 18 were mapped to the C1orf24 protein. The theoretical molecular weight of the most commonly sequenced splice form of C1orf24, containing both potential Met start codons is 72 kDa, consistent with the molecular weight estimate for the region of the 1D gel from where the protein was isolated (approximately 80 kDa). Thus the primary verification of a GENSCAN prediction was by proteomics, and the gene and protein structure finally elucidated by utilisation of genomic, proteomic and transcriptomic information.

With the substantial completion and publication of the human genome, it is expected that the nucleotide sequences for most, if not all, protein-coding genes will be contained in the sequence of the genome. For the various reasons discussed earlier, the goal of discovering the location and identity of such genes is likely to be achieved definitively via sequence analysis of expressed human proteins.

A completed description of these genes has utility in many areas of biology. Firstly, it organises the human

genes, has been achieved in a relatively short period. The full complement of protein-coding genes is readily completable by extending the number of cell and tissue types and protein separation methods to our proteomic and bioinformatic systems.

A further consequence of analysing a very large number of spectra in the manner described is that the same protein is observed in many tissues and disease states. The sample origin and molecular weight origin of the parent protein of every peptide from every gene is tracked throughout the pipeline and affords the opportunity to compare the observed proteins from a single gene in different backgrounds. This is analogous to EST sequencing and as with that approach we are able to observe alternative splicing as well as alternative protein processing (e.g. proteolysis) in generic and tissue specific settings.

The developments made in MS technology, protein separation, pre- and post- tandem mass spectrometry algorithms and bioinformatics have brought proteomics to a sensitivity and depth where it is contributing in a significant way to genomic annotation. We believe that the approach outlined here will become the definitive method by which protein coding genes and their structure, processing, and expression are described within the human genome.

ACKNOWLEDGEMENTS

The authors acknowledge the contribution made by Ensembl (<http://www.ensembl.org>) without which the work described herein would have been impossible. The authors acknowledge the encouragement and advice of R. A. Dwek, FRS. Tessella (www.tessella.com) is a software services company specialising in the support of scientific, technical and engineering establishments.

APPENDIX

During the completion of this work, the sequence of a human cDNA isolated from spleen and with significant similarity to this GENSCAN predicted gene (C1orf 24) was submitted to EMBL (accession AK074069, submitted 15 Feb 2002).

ABBREVIATIONS

EST	= Expressed sequence tag
RT-PCR	= Reverse transcription-polymerase chain reaction
1D gel	= One dimensional gel
2D gel	= Two dimensional gel
ICAT	= Isotope-coded affinity tagging

MTM	= Mass - trimer - mass
MALDI-TOF	= Matrix assisted laser desorption/ionization - time of flight
PTM	= Post-translational modification

REFERENCES

- Adam, P.J., Boyd, R., Tyson, K.L., Fletcher, G.C., Stamps, A., Hudson, L., Poyser, H.R., Redpath, N. *et al.* (2003). Comprehensive proteomic analysis of breast cancer cell membranes reveals unique proteins with potential roles in clinical cancer. *J. Biol. Chem.* **278**: 6482-9.
- Aparicio, S. A. J. R. (2001). How to count...human genes. *Nature Genet.* **25**: 129-30.
- Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R. *et al.* (2001). The Human Transcriptome Map: clustering of highly expressed genes in chromosomal domains. *Science* **291**: 1289-92.
- Choudhary, J.S., Blackstock, W.P., Creasy, D.M., and Cottrell, J.S. (2001). Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* **1**: 651-67.
- Eng, J.K., McCormack, A.L. and Yates III, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom* **5**: 976-89.
- Ensembl - a joint genome annotation project between the European Bioinformatics Institute (EBI) and the Sanger Institute.
- Ewing, B. and Green, P. (2000). Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25**: 232-34.
- Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H. and Aebersold R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.* **17**: 994-9.
- International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome *Nature* **409**: 860-921.
- Kuster, B., Mortensen, P., Andersen, J.S. and Mann, M. (2001). Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* **1**: 641-50.
- Liang, F., Holt, I., Perlea, G., Karamycheva, S., Salzberg, S.L. and Quackenbush, J. (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet.* **25**: 239-40.
- Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. *Nature Genet.* **30**: 13-19.
- Page, M.J., Amess, B., Townsend, R.R., Parekh, R., Herath, A., Brusten, L., Zvelebil, M.J., Stein, R.C. *et al.* (1999). Proteomic definition of normal human luminal and myoepithelial breast cells purified from reduction mammoplasties. *Proc. Natl. Acad. Sci. USA* **96**: 12589-94.
- Robinson, A.W., and Townsend, R.R. (2002). Automated identification of peptides. *International Patent Publication Number*: WO 02/21139 A2.
- Rohlf, C. and Southan, C. (2002). Proteomic approaches to central nervous system disorders. *Curr. Opin. Mol. Ther.* **4**: 251-8.
- Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A. *et al.* (2001). Experimental annotation of the human genome using microarray technology. *Nature* **409**: 922-7.
- de Souza, S.J., Camargo, A.A., Briones, M.R., Costa, F.F., Nagai, M.A., Verjovski-Almeida, S., Zago, M.A., Andrade, L.E. *et al.* (2000). Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proc. Natl. Acad. Sci. USA* **97**: 12690-3.
- Taylor, J. A. and Johnson, R.S. (1997). Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **11**: 1067-75.
- Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M. *et al.* (2001). The sequence of the human genome. *Science* **16**: 1304-51.