REVIEW

Last rolls of the yoyo: Assessing the human canonical protein count [version 1; referees: 1 approved, 2 approved with reservations]

Christopher Southan

IUPHAR/BPS Guide to Pharmacology, Centre for Integrative Physiology, University of Edinburgh, Edinburgh, EH8 9XD, UK

v1 **First published:** 07 Apr 2017, **6**:448 (doi: 10.12688/f1000research.11119.1)
Latest published: 07 Apr 2017, **6**:448 (doi: 10.12688/f1000research.11119.1)

Abstract

In 2004, when the protein estimate from the finished human genome was only 24,000, the surprise was compounded as reviewed estimates fell to 19,000 by 2014. However, variability in the total canonical protein counts (i.e. excluding alternative splice forms) of open reading frames (ORFs) in different annotation portals persists. This work assesses these differences and possible causes. A 16-year analysis of Ensembl and UniProtKB/Swiss-Prot shows convergence to a protein number of ~20,000. The former had shown some yo-yoing, but both have now plateaued. Nine major annotation portals, reviewed at the beginning of 2017, gave a spread of counts from 21,819 down to 18,891. The 4-way cross-reference concordance (within UniProt) between Ensembl, Swiss-Prot, Entrez Gene and the Human Gene Nomenclature Committee (HGNC) drops to 18,690, indicating methodological differences in protein definitions and experimental existence support between sources. The Swiss-Prot and neXtProt evidence criteria include mass spectrometry peptide verification and also cross-references for antibody detection from the Human Protein Atlas. Notwithstanding, hundreds of Swiss-Prot entries are classified as non-coding biotypes by HGNC. The only inference that protein numbers might still rise comes from numerous reports of small ORF (smORF) discovery. However, while there have been recent cases of protein verifications from previous miss-annotation of non-coding RNA, very few have passed the Swiss-Prot curation and genome annotation thresholds. The post-genomic era has seen both advances in data generation and improvements in the human reference assembly. Notwithstanding, current numbers, while persistently discordant, show that the earlier yo-yoing has largely ceased. Given the importance to biology and biomedicine of defining the canonical human proteome, the task will need more collaborative inter-source curation combined with broader and deeper experimental confirmation *in vivo* and *in vitro* of proteins predicted *in silico*. The eventual closure could be well be below ~19,000.

Open Peer Review

Referee Status: ✔ ? ?

	Invited Referees 1	2	3
version 1 published 07 Apr 2017	✔ report	? report	? report

1 **Michael Tress**, Spanish National Cancer Research Centre (CNIO) Spain

2 **Elspeth Bruford**, European Molecular Biology Laboratory UK

3 **Sylvain Poux**, Swiss Institute of Bioinformatics Switzerland, **Lionel Brueza**, Swiss Institute of Bioinformatics Switzerland

Discuss this article

Comments (0)

Corresponding author: Christopher Southan (cdsouthan@gmail.com)

How to cite this article: Southan C. **Last rolls of the yoyo: Assessing the human canonical protein count [version 1; referees: 1 approved, 2 approved with reservations]** *F1000Research* 2017, **6**:448 (doi: 10.12688/f1000research.11119.1)

Grant information: The author was supported for part of this work by the Wellcome Trust (grant number, 108420/Z/15/Z).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 07 Apr 2017, **6**:448 (doi: 10.12688/f1000research.11119.1)

Introduction

While hypothesis-neutral scientific endeavours are sometimes referred to in derogatory terms as "stamp collecting", the collation of molecular part lists (e.g. genes, transcripts, proteins and metabolites) remains a crucially important exercise, not only for many aspects of basic biology, but also for application to the biomedical sciences and drug discovery. Paradoxically, however, despite technical advances in analytical experimentation that should be making them easier to verify and quantify, definitive (or "closed") counts of even just these four entities for key species remain largely refractive. This is particularly so for proteins, as the most demonstrably biologically functional of these entity sets, even though they were the first to emerge historically by many decades[1]. In 2001, an analysis of the first public version of the draft human genome included an estimate of ~24,500 protein-coding genes[2]. The general opinion at that time was that this was lower than expected and would thus probably rise above 30,000. Notwithstanding, when the more complete first reference assembly (92% euchromatic coverage at 99.99% accuracy) was released in May 2004, the estimate was revised slightly downwards to ~24,000[3]. In the same year a detailed review appeared supporting a lower bound of ~25,000[4]. This latter publication alluded to a "yoyo" effect that persisted in subsequent reviews by falling to ~20,500 in 2007[5], rising to 22,333 in 2010[6], but then dropping to ~19,000 by 2014[7]. Those accepting the latter estimate may have felt a touch of chagrin as the count thereby fell to ~ 1000 below the model worm *Caenorhabditis elegans*. While we humans were still, reassuringly perhaps, ~ 7000 proteins ahead of the model fly *Drosophila melanogaster*, we are still ~20,000 behind the lowly *Paramecium* (see Table 1).

This article will compare and discuss the current numbers (as of 1Q 2017) from major sources. The evidence types and theory behind protein counting have been described in many publications and documentation from the individual database portals, but the reviews referenced above provide complementary background. It needs to be stated that numbers used herein refer to what can be termed the "canonical" human proteome. This has its origins in the Swiss-Prot approach to protein annotation whereby protein sequence differences arising from the same genomic locus either by alternative splicing or alternative initiations (or permutations of both) and/or genetic variants, are all cross referenced to a single, maximal length, protein entry[8]. Importantly, while this was originally introduced as the curatorial strategy of choosing the longest mRNA for an entry, it actually turns out to have post-genomic data support, not only in the form that coding-loci express a single main protein (i.e. that most predicted alternative transcripts may not be translated), but also that in most cases this is the max-exon form (i.e. the curatorial choice actually seems to be the biological "default")[9].

Historical growth

The set of open reading frames (ORFs) constituting the canonical human proteome can be historically followed in Ensembl and Swiss-Prot (as the manually reviewed and expert annotated sub-set of UniProtKB). Both of these are very different pipelines, but are partially coupled in the sense that the latter is one of the inputs to the automated ORF-building algorithms of the former. We can assess the progress of Ensembl first, since it has been compiling an approximation to the human proteome based on genomic predictions since 2001[10]. A 2004 review assessed historical figures from the first three years, over which the total shifted only marginally from 24,037 to 24,046[4]. While a maximum of 29,181 was reached in January 2002, this was an artefact associated with clone orientation changes caused by a switch in the assembly source, and this number had dropped back to 24,179

Table 1. Human protein coding gene counts from nine different portals, collected at the beginning of 2017. Ensembl numbers for the yeast, worm, fly and a protozoan are included for comparison (abbreviations are defined in the text).

Source	Version/date	Total	URL
GeneCards	v4.3.4, Jan 2017	21,819	http://www.genecards.org/
GeneID	Feb 2017	20,671	https://www.ncbi.nlm.nih.gov/gene/statistics/
Swiss-Prot	Release 2017_01	20,171	http://www.uniprot.org/
neXtprot	Jan 2017	20,159	https://www.nextprot.org/about/statistics
GENECODE	v25, March 2016	19,950	http://www.gencodegenes.org/stats/current.html
Ensembl	87.38	19,915	http://www.ensembl.org/Homo_sapiens/Info/Annotation
Vega/Havana	Feb 2017	19,768	http://vega.sanger.ac.uk/Homo_sapiens/Info/Annotation
HGNC	Feb 2017	19,033	http://www.genenames.org/cgi-bin/statistics
CCDS	20, Aug 2016	18,891	https://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi
S. cerevisiae	Dec 2011	6,692	http://www.ensembl.org/Saccharomyces_cerevisiae/Info/Annotation
C. elegans	WS250, 2012	20,362	http://www.ensembl.org/Caenorhabditis_elegans/Info/Annotation
D. melanogaster	Release 6, 2014	13,918	http://www.ensembl.org/Drosophila_melanogaster/Info/Annotation
P. tetraurelia	v87.1, 2006	39,642	http://protists.ensembl.org/Paramecium_tetraurelia/Info/Annotation/#assembly

by the next release. Despite some year gaps (not covered by the current archived data sets) the older figures can be plotted with the most recent ones to give a 15 year-span (Figure 1).

It is important to note that, for technical reasons, the longitudinal Ensembl protein numbers are not strictly comparable, since the pipeline model, its parameterisations and data feeds, have, as one might expect, evolved considerably over the years (e.g. the assembly source change mentioned above). This has included incremental improvements of various kinds (e.g. in the quality of the reference genome), but some changes have altered the exact definitions of the headline protein numbers. For example, the pseudogene figures given in the early 2001–3 releases needed to be subtracted from the totals. Those earlier numbers also specified a proportion of novel genes (defined as not having an exact match to RefSeq or UniProt entries at build time), but these tailed off from a maximum of 12,398 in November 2001 to only 46 by 2009 (release 54).

The most recent releases have other changes that complicate protein counts. One of these is the inclusion of "alternative sequence", referring to genomic sections that differ from the primary contiguous assembly. The current release of Ensembl (87.38) specifies 2,541 proteins in this category, but it is not clear which of these are just variants of those derived from the primary assembly. Another, somewhat enigmatic aspect, is the appearance in the protein count of so called "read-through" genes. These are defined as transcripts connecting two independent loci on the same strand. These debuted at 463 in release 74, via manual annotation, climbing slowly to the current total of 526. While they are also included in the NCBI genome annotation, these have not been included in the Figure 1 counts because, if they exist at all as translated chimeric proteins, they are non-canonical by definition.

Despite these shifts in exactly what the protein numbers represent, we can draw three principle conclusions from Figure 1. These are: a) yo-yoing has at least subsided, if not ceased; b) the number has plateaued at just below 20,000; and c) the pipeline has ceased to spawn significant numbers of novel proteins (i.e. they are now predominantly "seen before").

One of the core operations for Ensembl is resolving transcripts and their mRNA coding sections (CDSs) against ORFs predicted *ab initio*. Swiss-Prot, on the other hand, has historically been doing this for mRNA-to-protein independently of genomic coordinates (although it increasingly now maps the two together where possible). Over the years, the criteria and manual triage for defining canonical ORFs have been consistently applied in Swiss-Prot. This means the growth rate can be straightforwardly recorded by slicing Swiss-Prot human proteins by "create date" (Figure 2). The pattern is interpretable as a concerted effort towards provisional closure of the proteome at 19,658 by 2008. Subsequent increases were essentially incremental, climbing slowly to 20,168 by 2017.

While issues around evidence types will be addressed later, a simple filter can be applied to count just those proteins with either transcript and/or other forms of experimental support for their existence. The result, in the Figure 2 plot, shows this difference to be fairly constant (i.e. that in the order of ~1,400 sequences remaining experimentally unsupported). There are three other salient features. The first is that the total has only increased by a modest 516 since 2009, whereas Ensembl shrunk by 1,455 over the same period. They have thus both converged towards ~20,000 (it is not clear if the two sets are congruent for the same ORFs, but this question will be addressed later). However, there were already

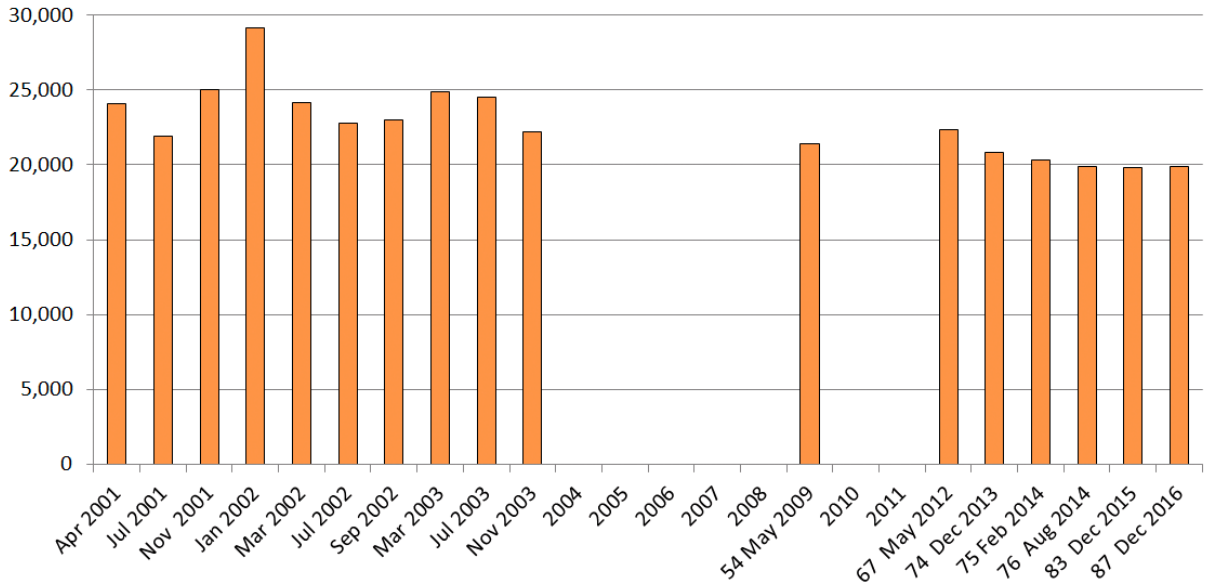

Figure 1. Protein counts from the Ensembl pipeline database releases over the first three years and last seven years. The latter are only those from the current archive that have protein rebuilds rather than maintenance/patch releases with nearly identical numbers.

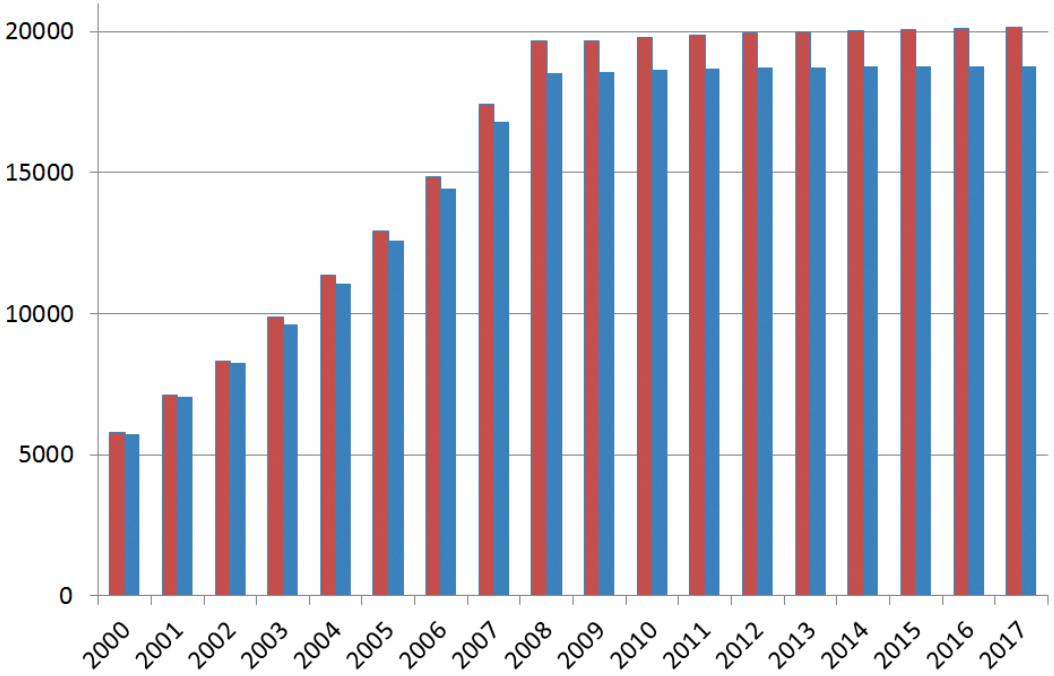

Figure 2. Protein counts from human UniProtKB/Swiss-Prot by create date (red). The blue columns include the additional selection for existence evidence at the protein or transcript levels (note the date is just for the entry into Swiss-Prot, not the first appearance of the sequence in TrEMBL that can be many years earlier).

indications of approximate concordance as early as 2001, where adding the Ensembl novels to the Swiss-Prot knowns reached 18,191. The inference is that the number of novel proteins confirmed since 2001 is less than 2000. Note also that many are TrEMBL-to-Swiss-Prot promotions (i.e. with data already surfaced) rather than *de-novo* deposited protein sequences. By comparing 2009 with the subsequent seven years we can also infer that Swiss-Prot has not purged significant numbers of accessions (i.e. they have revised sequences but generally not removed them).

Current counts

We can move on from tracking historical numbers to taking a contemporary snapshot of major sources (including the two already described) that are well established and regularly declare revised protein counts (Table 1). There are many aspects that could be expanded on from this set, but the feature that immediately stands out is the difference of nearly 3000 between highest and lowest (i.e. 13%). The highest figure comes from what can be considered a meta-source, GeneCards, that merges different pipeline outputs, so this could be expected to be an upper bound[11]. The protein-coding set from the NCBI genome annotation pipeline ranks second but there are some caveats regarding comparability with the other sources[12]. One of these is the inclusion of 1235 "LOC" entries with low homology support. Although 107 of these do have Ensembl gene IDs, none have been assigned Human Gene Nomenclature Committee (HGNC) symbols. Removing LOCs from the NCBI protein set would drop them down to seventh at 19,436.

The next two sources are related in that neXtprot takes the human Swiss-Prot set as a starting point for evidence expansion and interrogation enhancements. This is why these have (almost) the same count (the residual differences being due to synchronisation timings)[13]. The next three sources are also coupled in the sense that not only are GENECODE and Vega marked-up in Ensembl, but there are plans to merge the three. However, they do show a small difference of 182, with the lowest being the Vega pipeline (as Havanna manual curation). But even from Vega, there is a substantial drop of 735 to the stringently reviewed approved protein-coding gene-based assignments from the HGNC. The lowest number in Table 1, coming in at just below 19,000, comes from the Consensus Coding Sequence (CCDS) project. These correspond to a core set of proteins annotated as having full length transcripts that exactly match reference genome coordinates.

Some sources have invested effort into mapping between each other's identifiers. This can establish if the protein sequence in pipeline output A is the same as pipeline B. However, the fidelity of such a mapping (and consequent cross-reference reciprocity) depends on differences in methods and stringencies. For example, for all intents and purposes the beta-secretase 1 entry (BACE1) is the same across all 9 pipelines. However, a different population variant was chosen on each side of the Atlantic. Therefore, the RefSeq and Gene ID sequence NP_036236 differs by one residue (481 Cys → Arg) from the Swiss-Prot and Ensembl sequence as P56817. Note also that HNGC does not instantiate sequence entries in the way that the other pipelines do, but

collates cross-references, so in this case HGNC 933/BACE1 points to both sequences. The process of cross-referencing between multiple annotation sources allows the generation of both intersects and differences. Crucially, in terms of protein counting, this gives us the possibility to discern where they are concordant or discordant and (on a good day) we may be able to identify causes for the latter.

Cross-reference counting

All nine sources in Table 1 provide some extent of cross-referencing between what should be the same protein in different sources (also referred to as cross-mapping). However, the choice was made here to exemplify just four identifiers, Swiss-Prot accession numbers, HGNC IDs (directly, or via the current gene symbols) Ensembl gene IDs and NCBI Entrez Gene IDs. These were chosen for their global prominence but also methodological complementarity. This derives from the fact that that the first two are essentially automated pipelines (but different), while the second two are primarily manual expert annotation operations (but also different). Each of the four offers their own internal ways of querying cross-references, including BioMart installations[14] or downloadable mapping tables for this to be done extrinsically. However, because it has the largest number of selectable cross-references, as well as extended options for live-linked result displays and filtered downloading, the UniProt interface was used here. Intersects for the four sources can be seen in Figure 3.

Figure 3 can be explained as follows: The queries executed gave the totals indicated in the segments. Note that some segments are empty, because, by definition, the identifier mapping has been done "inside" Swiss-Prot (even if in some cases the external sources collaborated in generating the mappings). By comparing with Table 1, we can thus see that 2,923 NCBI proteins did not map at all (which includes most of the LOCs). Similarly, 834 Ensemble protein gene IDs also did not map. For HGNC, on the other hand, we see the cross-reference result is actually 905 higher than the distinct identifier count at source. One explanation could be a proportion of a one-to-many relationship (e.g. Swiss-Prots with more than one HGNC). Some were identified, such as haemoglobin subunit alpha (P69905) that maps to HGNC HBA1 and HBA2.

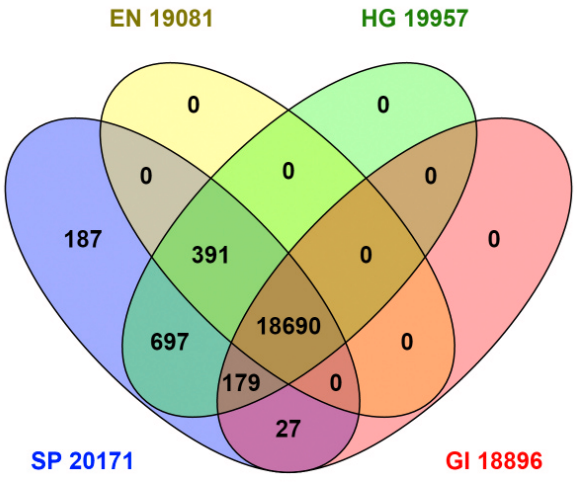

Figure 3. Intersects between identifier cross-references recorded from the UniProt interface. The results are generated via cross-reference totals according to UniProt, not from the sources *in situ*. EN, Ensembl; SP, Swiss-Prot; HG, Human Gene Nomenclature Committee; GI, NCBI Entrez Gene.

A notable result from Figure 3 is that a 1:1:1:1 mapping (i.e. four-way concordance) is achieved for only 18,690 proteins, lower than any of the totals from Table 1. Detailed analysis of all the segments cannot be presented here but some trends can be noted. Starting with the 187 in the "SP" segment (i.e. Swiss-Prot only, absent from the other three), the majority of the protein names are given as "putative" or "uncharacterised". The 391 common elements in "SP", "EN" and "HG" (i.e. missing in NCBI Gene) are clearly dominated by variable domains of immunoglobulin light chains and HLA class I histocompatibility antigen alpha chains, the polymorphic nature of which necessitates a level of manual annotation that may not have been compatible with the NCBI pipeline automation. The 179 common elements in "SP", "HG" and "GI" (i.e. missing in Ensembl) are enriched for "Uncharacterized protein" from the so called Chromosome ORF predictions. The large set of 697 common elements in "SP" and "HG" (i.e. missing in NCBI Gene and Ensembl) are heterogeneous but show enrichment for translated endogenous retrovirus transcripts, putative uncharacterized proteins encoded by LINC loci and include 41 odour receptors. Notably, in these three sets, the HGNC cross-references classify them as not being within their own protein-coding set of 19,033, but rather as endogenous retrovirus, long non-coding RNAs and pseudogenes, respectively. This particular discordance (i.e. in UniProt but not a protein according to the HGNC) explains the 1: many cross-references mentioned at the start of this section. A duplicate check on the 960 indicated only 152 could be ascribed to Swiss-Prots with multiple HGNCs. It can also be seen in Figure 3 that two of the Swiss-Prot intersects are empty. The explanation is that Ensemble and NCBI Gene have consolidated mapping reciprocity for proteins in Swiss-Prot (but, as mentioned above, many proteins from these two sources are still nominally "outside" Swiss-Prot).

As one of its powerful utilities, we can interrogate ~ 90 cross references in UniProt. While not all of these are human-relevant we can chose those to compare with Table 1. This has already been done for the four above but can be extended. For example, we can determine counts of 18,384 from CCDS and 19,940 for GeneCards. Note both of these are below the *in situ* counts by 510 and 1,871 respectively (GENCODE and Vega do not currently have cross-references inside Swiss-Prot). In some cases it may be possible to investigate counts reciprocally. For example, from the HGNC protein-coding download table we can establish that the 19,035 rows in the UniProt mapping column contained 18,997 Swiss-Prot IDs. The same table includes 19,035 Vega Gene ID mappings that also collapse to 18,973 distinct entries. This confirms what was implicated already above, as a small proportion of multiple Swiss-Prots < > HGNCs is also occurring for HGNC < > Vega. Cross-mapping counts can similarly be explored via other sources for comparison, depending on what query and/or download options are available. However, accumulating such results can

quickly generate large Venn-type sets that generally end up being more confusing than illuminating.

Following on from above, since they are derived from structure data sources, cross-references give precise protein counts; but they also have associated equivocality (even though they will be used further in this report). For this reason, it is important to understand (e.g. via source documentation) technical differences in exactly how the mappings are determined. A second problem is they may be circular (i.e. source B may collegially accept A< > B mappings from source A without independently verifying the reciprocity of B >A). The third problem is synchronisation, where release dates are at different intervals (and may not always include mapping refreshes). The forth problem is the "churn" rate (appearance and/or disappearance of protein records) in genome resources. This is much lower that is was some years ago, but can still be an issue.

Existence evidence

In the context of advancing towards proteomic "closure", the imperative to verify the existence of an *in silico* database ORF as an *in vivo* protein translation product is obvious. By definition, the prerequisite mRNA transcription also needs experimental verification; especially if the ORF is only a genomic DNA prediction. However, on its own, active transcription is insufficient to prove translation, even with a predicted CDS, and it is established that pseudogenes can exhibit low-level transcription[15]. While it has inherited the categorisations from UniProt, the neXtprot database has a particular focus on the evidence code system and has set up collaborations to extend experimental support in general[13]. The outlines of this can be seen in Figure 4.

The categories (expanded on in the neXprot documentation) are as follows:

1. PE1: evidence that includes at least partial Edman sequencing, mass spectrometry (MS) with a threshold of 2 peptides of at least 9 amino-acids, X-ray or NMR structure, protein-protein interaction data or detection by antibodies (Abs).
2. PE2: not proven at protein level but has transcription data (e.g. cDNA, RT-PCR or Northern blots).
3. PE3: probable existence based on orthologues with high similarity scores being found in related species.
4. PE4: no evidence at the protein, transcript, or homology levels.
5. PE5: may be a spurious *in silico* translation of a non-coding transcript.

There is now a community effort to promote more proteins to P1 using both MS and Abs, so we can go into these in more detail. The former has a long history with a proprietary project reporting MS identification of 14,223 human proteins as early as 2004[16]. An analogous public effort described the verification of 11,115 Ensembl coding sequences, made available in the first data release of the ProteinAtlas (PA) in 2005[17]. By 2017 the Human Proteome Organisation has been extensively engaged in MS initiatives, particularly in regarded to the "missing proteins" (i.e. those still in P2 to P5) that remain refractory to tryptic peptide verification at the necessary stringency. This aspect has been the subject of several recent reviews and so does not need expanding here[18,19].

As another important methodological push, antibody-based proteomics has developed more recently into a large-scale enterprise. This was first described in 2014 as the Human Protein Atlas project with its own associated database[20]. This has now been extended with the setting up of an International Working Group for Antibody Validation and the accompanying Antibodypedia database[21]. These have the objective to increase the reproducibility of protein identification and ultimately, as with the MS initiatives, to move more sequences up to the P1 evidence code)

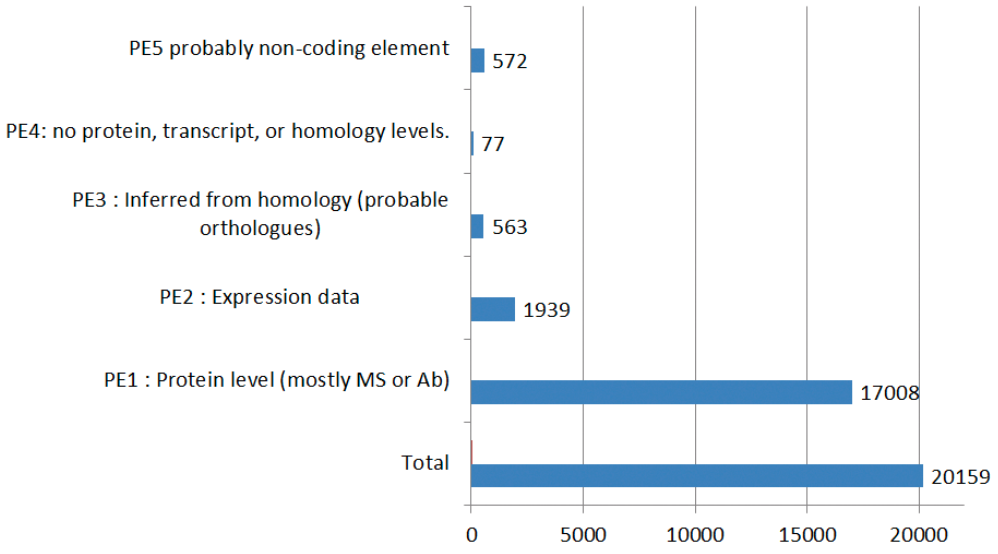

Figure 4. Protein existence codes and their occupancy statistics from neXtprot.

We can use the categories above to further "slice and dice" cross-referencing to gain more insight into particular subsets (e.g. via downloadable identifier sets for P1to P5). The possible query combinations are many, so we need to frame useful questions. Notably, it is now possible to select proteins supported by PA MS support entries (17,084) or HPA (16,800) or both (15,189) (n.b. numbers differ slightly from those in neXtprot of 18,083 for PA and 16,473 for HPA). In terms of questions, an example that can be posed is "how many proteins, either supported by HPA or PA, overlap with the 4-database consensus set generated in Figure 3?" The result (Figure 5) effectively intersects the *in silico* with the in *vivo* evidence sets.

As was done for Figure 3, lists from the Venn sections were input to the UniProt ID mapping interface to examine trends. Not all of these can be discussed here, but looking at the unique sets exposed some initially counter-intuitive results. For example, the 4-way only (734) included 214 P1s, but without HPA or PA cross-references. This is because P1 also includes 3D structures and interaction data. Looking at the 152 HPA-only set included 101 at P4 or P5 levels (i.e. unexpectedly high for the implied Ab confirmation which might be expected to push them up to P1). It turns out there is a cross-reference specificity problem from the inclusion of uncertain results. The HPA link (for the 16,800) actually means the protein has been tested (i.e. had an antibody raised against peptide sections) but is not necessarily confirmed. The histochemistry support status, including consistency with two sources of transcript data are commented on in each HPA entry. However, from the HPA download for 16.1, only 10,230 (of the Ensembl proteins as primary identifier) are designated as "approved" or "supported" at the histochemistry level. Examples of evidence complications include the 40-residue of putative protein FAM86JP as the Swiss-Prot entry Q05BU3. Flagged as P5, this shows anomalies including designation as a pseudogene by HGNC (n.b. it has neither GeneID nor an Ensembl cross-reference which excluded it from the 4-way set) and the HPA entry ENSG00000186523-FAM86B1 was flagged as uncertain based on two antibodies. A second example exposes a different problem. The putative uncharacterized protein C7orf76 (Q6ZVN7) is mapped from UniProt to a different protein in HPA as ENSG00000127922-SHFM1 (i.e. P60896). The miss-mapping appears to be extrinsic to HPA and in this case could be a UniProt < > Ensembl problem (which is why this is not in the 4-way set). It is important to emphasise that none of this is about fault finding, but these examples attest to the technical challenges of evidence classifications and mapping fidelity.

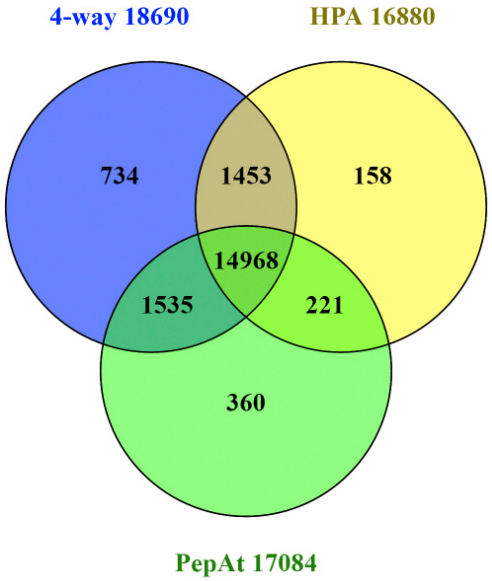

Figure 5. Swiss-Prot cross-reference intersects between the 4-way confirmed set from Figure 3, the Human Protein Atlas (HPA), and the Peptide Atlas.

Inspecting the 360 "PepAt" (i.e. PeptideAtlas only) set reveals a different set of interpretive challenges. An example is the smallest of the set at only 11 residues as morphogenetic neuropeptide (P69208). This has no genomic annotation, but does have an apparent match in PeptideAtlas for the peptide QPPGGSKVILF. The Swiss-Prot entry has its origins in an Edman sequencing result from 1986 and is consequently indicated as "Experimental evidence at protein level", but has been dropped from neXtprot. A large proportion of the rest of the 360 are immunoglobulin heavy variable and HLA class I histocompatibility antigen chains for which the ability of the PeptideAtlas system to resolve into separate proteins is unclear.

Small proteins

Back in 2004, it was already mooted that a significant expansion in protein number was likely to occur via the discovery of small ORFs (smORFs). However, this was not supported by Swiss-Prot statistics at that time[4]. In the intervening decade, the smORF question has surfaced regularly[22] and it now overlaps with the two closely related themes of *de novo* protein evolution (i.e. recent non-coding to coding transitions)[23] and ribosomal profiling experiments attempting to define the translation of novel smORFs from what was hitherto classified as non-coding RNA[24]. In addition, the theme of existence evidence discussed above is also relevant, since whatever data support type is being sought (e.g. active transcription plus detection by MS or Abs), the experimental verification of smORFs becomes more difficult.

An obvious approach to this topic is to repeat the exercise first performed in 2004[4], namely splitting the smORF count in Swiss-Prot by create date. By setting a cut-off of 100 residues, the current total is 682/20,168. This can be compared with the corresponding 2009 totals of 612/19,675. This establishes that the proportional smORF content has only risen from 3.1% to 3.3%. In addition, from the latest 2017 size cut, 161 of the 682 do not have an HGNC biotype designation as protein-coding. Many also only have the protein existence support as Edman sequencing reads from the earliest Swiss-Prot releases. These short sequences are difficult to genome map and/or re-confirm by MS, which is why six were recently purged from neXtprot (P.Gaudet personal communication). We are thus presented with a paradox that, despite many reports of putative novel human smORF discovery, very few are crossing the Swiss-Prot evidence threshold for becoming new protein entries.

Notwithstanding, recently confirmed smORF examples have surfaced that are informative from the protein counting viewpoint. The first of these, the apelin receptor early endogenous ligand, was integrated into Swiss-Prot in 2014 (HGNC symbol APLEA, synonyms Elabela, Toddler; see Swiss-Prot P0DMC3 for cross-references including links to the discovery papers). It was in fact "hiding in plain sight" in so far as its full-length cDNA (AK092578) had been in GenBank since 2008. However, since this sequence translates into eight possible smORFs, the submission process for the high-throughput cloning project (sensibly) chose not to annotate a CDS in the feature lines of this prostate library entry, since there was no basis on which to choose any of the possible translations by protein similarity at that time (although arguably, manual sequence analysis, including TBLASTX, might have given clues). Significantly though, this transcript had originally been annotated in Vega as a Long non-coding RNA (LncRNA) giving rise to speculation that additional cryptic smORFs could be "hiding" in other LncRNAs. Such a second case has in fact been described in 2016 in paper entitled "A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle", although the work was done in mouse[25]. The publication was processed by Swiss-Prot in March 2016 to generate P0DN83 and P0DN84 for a 34 residue mouse and human proteins, respectively.

These two smORFs illustrate a spectrum of evidence differences as follows:

- In terms of transcript support, APLEA has been re-cloned as KJ158076 with a submitted CDS, but this is not yet incorporated in the Swiss-Prot annotation. The DWORF authors mention obtaining cDNAs but have neither deposited human or mouse mRNA accession numbers. There are many TBLASTN matches as supporting evidence for the protein (not withstanding miss-matches, see below) both to mammalian sequences designated as LOC non-coding RNAs and over 30 human expressed sequence tag (EST) mRNAs.
- APLEA has three-way genomic support and a CCDS, while DWORF has no human genome cross reference in Swiss-Prot. The mouse paralogue does have an Ensembl protein mapping (ENSMUSG00000103476) despite still being flagged as an LncRNA gene in the Mouse Genome Atlas. However, multiple lines of evidence (Southan, unpublished observations) indicate the correct human sequence is the 35 resides represented in ENSG00000240045 (via Vega) as TrEMBL A0A1B0GTW0 (but circularly as this was picked up from Ensembl) and independently as ACT64388 from 2009. The predicted transcript is classified by NCBI as a non-coding LOC100507537.
- Neither APLEA nor DWORF have any cross-references in the seven MS sources in UniProt. Note that APLEA cannot pass the double 9-mer criteria for neXtprot, and DWORF only has a single predicted tryptic peptide. Whether either protein passes the verification threshold for MS datasets in the future remains to be seen.
- Publications for both APELA and DWORF have included Western blots from Abs raised against peptides (but mouse for the latter). However, neither yet has an HPA entry. While the possibility of inclusion in a future update is clear for APELA, there may not only be technical challenges from the small size of DWORF, but also, since HPA uses Ensembl IDs for its primary identifiers, this protein and its transcript would need first to be resolved in a future Ensembl release (n.b. LOC100507537 appears to have somehow parsed HPA transcript data, but this may be a miss-mapping).
- Replication of the basic findings and expanded aspects of *in vivo* function have been consolidated in numerous publications for APELA, including a 2017 paper[26]. While the experimental characterisation of DWORF rests on one study done with mouse so far[25], consolidation of the human protein evidence is to be expected in forthcoming work.

To summarise the implications; the discovery of additional smORFs seems certain, especially given that the putative LncRNA gene count has recently risen to 27,919[27]. However, the question remains as to how many will be verified to the evidence level sufficient to enter the major genome and protein portals (even though it will be challenging to obtain Abs and MS verification data). On a continuum of what we might expect between 10, 100 or 1000, the middle estimate seems most likely.

Pharmacological interaction intersects

This last section assesses the corroboration of data linkages by existence evidence and other types of concordance. Many of the Swiss-Prot cross-references are related to protein function and other attributes such as tissue distribution or post-translational modification. Others would include pathway membership, protein-protein interactions, Genome Ontology categorisation, disease associations, interactions between enzymes and substrates, drugs and their targets, as well as endogenous ligands for receptor proteins. The advantage of the analyses described above is that results centred on functional categories can be intersected with independent cross-references. This can be exemplified by selecting the curated ligand interactions in the IUPHAR/BPS Guide to PHARMACOLOGY[28] (GtoPdb) that are included in the set of five chemistry (interaction) cross-references. The current UniProt has 1,460 human Swiss-Prot records (as defined by the GtoPdb criteria for submitting the links) that have publication-supported molecular interactions. The majority are pharmacologically active small-molecules, but the curated relationships include some protein-protein interactions, for example, antibody ligands directed against cytokine targets (n.b. a proportion of these proteins are derived from a new project as the Guide to Immunopharmacology). The result of the corroboration analysis is shown in Figure 6.

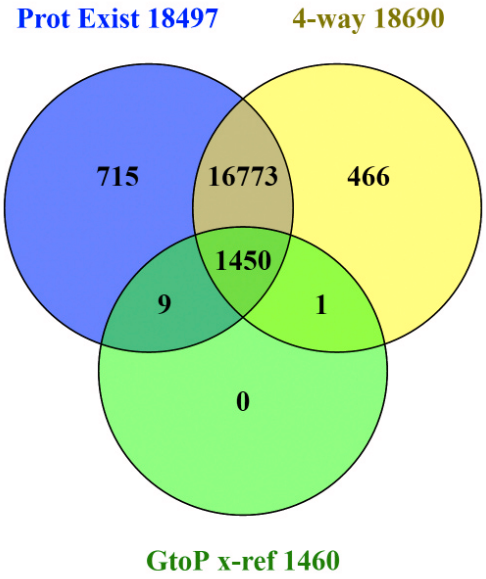

Figure 6. Corroboration of human proteins in Swiss-Prot with ligand interactions in GtoPdb (selected as "Guideto-PHARMACOLOGY" in the Chemistry cross-references). The first of the two intersected lists are labelled as "Prot Exist" with evidence at the transcript and protein levels (i.e. PE1 and PE2 from Figure 4), and the 4-way major source consensus set (i.e. the central panel of Figure 3).

We can see the results of a three way comparison in Figure 6 between existence evidence, four-source convergence and GtoPdb entries. The first feature to note is that not all proteins with existence evidence are in the four-source set, and vice versa. Possible systematic reasons behind this cannot be explored here, but may be related to the points discussed for Figure 5. The key observation for GtoPdb is that, reassuringly, 1,450 entries intersect with both existence evidence and four-source identifiers. Notwithstanding, there are nine intersects with the existence set but not four-source corroborated with one vice versa (i.e. in the four-source set but not evidence-supported). Given that GtoPdb interactions are expert-curated, the result from Figure 6 raises questions about the annotation of the 10 protein entries. These were followed up to establish that the lack of evidence support for P0C264 arises from the absence of an mRNA entry (i.e. it remains a genomic prediction). The existence of this kinase seems well supported (e.g. via CCDS74457), but a cloned cDNA would be an important consolidation. Inspection of the other nine sequences also supported their existence but they all had a mixture of cross-referencing failures that had excluded them from the four-source set. For example, for the aspartyl aminopeptidase, DNPEP (Q9ULA0) the protein is solidly supported even to the extent of a PDB structure, but the Entrez GeneID is missing (although this is cross-referenced by HGNC). Likewise, the alpha-2B adrenergic receptor, ADRA2B (P18089) is solidly supported, but in this was missing the Ensembl cross-reference (it turns out from Swiss-Prot update enquiry this was due to an unusual accession number change associated with a TrEMBL to Swiss-Prot transition, Gasteiger, personal communication). In both cases GtoPdb had in fact been manually curated in the correct links for the Entrez Gene ID in Target ID 1559 and Ensembl Gene for Target ID 26, respectively (n.b. the appropriate UniProt corrections have been suggested via the feedback form). This cross-checking for GtoPdb targets thus proved a useful exercise that will be re-visited as our protein content expands.

Conclusions

Despite over 16 years having elapsed since the first draft human genome, the diversity of current counts indicates that progress towards what the community might consider a gold-standard set of canonical protein sequences, remains frustratingly slow. This is especially so considering that the "zone of equivocality" lies only between an upper bound of ~ 20,000 and a lower one of ~18,500. The slow progress towards closure is clearly a reflection of both the inherent biological complexity of protein translation, as well as the challenges of combining automated annotation with various proportions of expert curation needed to define the entire expressed genomic landscape[29]. There are of course caveats, even with the concept of closure, in so far as recent evidence indicates that each of us, on average have at least 100 protein loss-of function variants (i.e. proteomes are "personal")[30].

The wider bioscience community could be forgiven being puzzled that major global efforts continue to produce different sets of canonical proteins at roughly the same time from the same primary data (leaving aside another layer of yet more inter-source differences in alternative splice and/or initiation forms). Those of us with some insight into the bioinformatic, genomic and proteomic challenges might be more sanguine in our judgment, but the criticism still stands (note also that human is the testbed from which the community needs progress to analogous proteomic closure for at least mouse, rat and Zebrafish). Approaching the question as to why this situation persists and possible solutions, would necessitate a detailed comparison of the underlying assumptions, data processing models and pipeline parameterisations. However, inter-source clustering of explicit protein sequences could make identifying difference more effectively than cross-references alone (e.g. a possible resurrection of the Human Protein Index initiative[31]).

Regardless of the technical options to solving the problem, substantial resources have been committed over decades by the major gene and protein annotation resources globally. We should thus expect more inter-team collaboration dedicated to harmonising amongst themselves for the mere ~2000 protein sequences in question (i.e. not many compared to the 0.55 million and 77 million processed in Swiss-Prot and TrEMBL respectively). It could be argued that additional (collective) manual curation would

be needed to accomplish this, but the consequent improvement *in silico* concordance could then be consolidated by an expansion of experimental existence verification both *in vitro* and *in vivo*. This could include a supply of expressed protein standards, advances in MS-based proteomics, including sets of synthetic proteotypic peptides for spiking experiments[32], deep transcript profiling by RNA-seq and the increased availability of validated antibody reagents.

Data availability

These statistics on protein numbers are presented and compared here in good faith and with implicit expectation that they should be reproducible, including by others who may want to repeat and/or extend these types of analyses. Notwithstanding, this may be confounded by several factors that could give rise to slightly different results (but it is hoped not major discrepancies). The most obvious is data updates that can be as frequently as monthly for some sources (e.g. since the completion of this work UniProt notched up to UniProt release 2017_03 on March 15, 2017 with the human SwissProt count increasing, from Table 1, by 13 proteins to 20,184). Another is the exact form of the queries, which vary between resources, particularly when each selection interface has a different look and feel, different syntactic formats of execution and download lists having different formats of cross-referenced identifier columns. One example is the need to covert UniProt interface queries into the equivalent SPARQL queries in neXtprot as shown below. The UniProt syntax to count HGNC cross-references, as entered in the web query box, is below:

database:(type:hgnc) AND reviewed:yes AND organism:"Homo sapiens (Human) [9606]"

The answer was 19967 (March 2017), but note we need to make to pre-selects for a) species/organism and b) "reviewed" to select Swiss-Prot over TrEMBL. For the neXtProt equivalent cross-reference query, these two pre-selects are not necessary since is human Swiss-Prot derived anyway. The HGNC select has the form below:

```
select distinct ?entry where {
 ?entry :reference ?ref .
 ?ref :provenance db:HGNC ;
  :accession ?ac.
 filter (regex(?ac,'^HGNC'))
}
```

In this case the result was 19956. The basic listings from sources used and some of the result sets have been made available as a Figshare data collection (https://figshare.com/collections/Supplementary_data_for_assessing_the_human_canonical_protein_count/3716413[33]). If any reproducibility issues do arise, interested parties are welcome to contact the author.

Competing interests

No competing interests were disclosed.

Grant information

The author was supported for part of this work by the Wellcome Trust (grant number, 108420/Z/15/Z).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The efforts of all the genomic and protein annotation teams referred to in this work are much appreciated. Discussions of discordances and other quirks should not be misinterpreted as criticism of the resources concerned. Thanks are due to those who answered questions on this topic on BioStars, various database helpdesks and Twitter, as well as Dr Pascale Gaudet for help with neXtprot queries.

References

1. Sanger F: **The arrangement of amino acids in proteins.** *Adv Protein Chem.* 1952; **7**: 1–67.
PubMed Abstract | Publisher Full Text
2. Lander ES, Linton LM, Birren B, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature.* 2001; **409**(6822): 860–921.
PubMed Abstract | Publisher Full Text
3. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature.* 2004; **431**(7011): 931–945.
PubMed Abstract | Publisher Full Text
4. Southan C: **Has the yo-yo stopped? An assessment of human protein-coding gene number.** *Proteomics.* 2004; **4**(6): 1712–1726.
PubMed Abstract | Publisher Full Text
5. Clamp M, Fry B, Kamal M, *et al.*: **Distinguishing protein-coding and noncoding genes in the human genome.** *Proc Natl Acad Sci U S A.* 2007; **104**(49): 19428–19433.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Pertea M, Salzberg SL: **Between a chicken and a grape: estimating the number of human genes.** *Genome Biol.* 2010; **11**(5): 206.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Ezkurdia I, Juan D, Rodriguez JM, *et al.*: **Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes.** *Hum Mol Genet.* 2014; **23**(22): 5866–5878.
PubMed Abstract | Publisher Full Text | Free Full Text
8. The UniProt Consortium: **UniProt: the universal protein knowledgebase.** *Nucleic Acids Res.* 2017; **45**(D1): D158–D169.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Tress ML, Abascal F, Valencia A: **Alternative Splicing May Not Be the Key to Proteome Complexity.** *Trends Biochem Sci.* 2017; **42**(2): 98–110.
PubMed Abstract | Publisher Full Text
10. Aken BL, Achuthan P, Akanni W, *et al.*: **Ensembl 2017.** *Nucleic Acids Res.* 2017; **45**(D1): D635–D642.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Fishilevich S, Zimmerman S, Kohn A, *et al.*: **Genic insights from integrated human proteomics in GeneCards.** *Database (Oxford).* 2016; **2016**: pii: baw030.
PubMed Abstract | Publisher Full Text | Free Full Text
12. NCBI Resource Coordinators: **Database Resources of the National Center for Biotechnology Information.** *Nucleic Acids Res.* 2017; **45**(D1): D12–D17.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Gaudet P, Michel PA, Zahn-Zabal M, *et al.*: **The neXtProt knowledgebase on human proteins: 2017 update.** *Nucleic Acids Res.* 2017; **45**(D1): D177–D182.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Smedley D, Haider S, Durinck S, *et al.*: **The BioMart community portal: an innovative alternative to large, centralized data repositories.** *Nucleic Acids Res.* 2015; **43**(W1): W589–W598.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Guo X, Lin M, Rockowitz S, *et al.*: **Characterization of Human Pseudogene-Derived Non-Coding RNAs for Functional Potential.** *PLoS One.* 2014; **9**(4): e93972.
PubMed Abstract | Publisher Full Text | Free Full Text

16. McGowan SJ, Terrett J, Brown CG, *et al.*: **Annotation of the human genome by high-throughput sequence analysis of naturally occurring proteins.** *Curr Proteomics.* 2004; **1**(1): 41–48.
Publisher Full Text

17. Desiere F, Deutsch EW, King NL, *et al.*: **The PeptideAtlas project.** *Nucleic Acids Res.* 2006; **34**(Database issue): D655–8.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Omenn GS, Lane L, Lundberg EK, *et al.*: **Metrics for the Human Proteome Project 2016: Progress on Identifying and Characterizing the Human Proteome, Including Post-Translational Modifications.** *J Proteome Res.* 2016; **15**(11): 3951–3960.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Segura V, Garin-Muga A, Guruceaga E, *et al.*: **Progress and pitfalls in finding the 'missing proteins' from the human proteome map.** *Expert Rev Proteomics.* 2017; **14**(1): 9–14.
PubMed Abstract | Publisher Full Text

20. Fagerberg L, Hallström BM, Oksvold P, *et al.*: **Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics.** *Mol Cell Proteomics.* 2014; **13**(2): 397–406.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Uhlen M, Bandrowski A, Carr S, *et al.*: **A proposal for validation of antibodies.** *Nat Methods.* 2016; **13**(10): 823–7.
PubMed Abstract | Publisher Full Text

22. Pueyo JI, Magny EG, Couso JP: **New Peptides Under the s(ORF)ace of the Genome.** *Trends Biochem Sci.* 2016; **41**(8): 665–678.
PubMed Abstract | Publisher Full Text

23. Schmitz JF, Bornberg-Bauer E: **Fact or fiction: updates on how protein-coding genes might emerge *de novo* from previously non-coding DNA [version 1; referees: 3 approved].** *F1000Res.* 2017; **6**(F1000 Faculty Rev): 57.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Mumtaz MA, Couso JP: **Ribosomal profiling adds new coding sequences to the proteome.** *Biochem Soc Trans.* 2015; **43**(6): 1271–1276.
PubMed Abstract | Publisher Full Text

25. Nelson BR, Makarewich CA, Anderson DM, *et al.*: **A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle.** *Science.* 2016; **351**(6270): 271–5.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Yang P, Read C, Kuc RE, *et al.*: **Elabela/Toddler Is an Endogenous Agonist of the Apelin APJ Receptor in the Adult Cardiovascular System, and Exogenous Administration of the Peptide Compensates for the Downregulation of its Expression in Pulmonary Arterial Hypertension.** *Circulation.* 2017; **135**(12): 1160–1173.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Hon CC, Ramilowski JA, Harshbarger J, *et al.*: **An atlas of human long non-coding RNAs with accurate 5′ ends.** *Nature.* 2017; **543**(7644): 199–204.
PubMed Abstract | Publisher Full Text

28. Southan C, Sharman JL, Benson HE, *et al.*: **The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands.** *Nucleic Acids Res.* 2016; **44**(D1): D1054–68.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Mudge JM, Harrow J: **The state of play in higher eukaryote gene annotation.** *Nat Rev Genet.* 2016; **17**(12): 758–772.
PubMed Abstract | Publisher Full Text

30. Narasimhan VM, Hunt KA, Mason D, *et al.*: **Health and population effects of rare gene knockouts in adult humans with related parents.** *Science.* 2016; **352**(6284): 474–477.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Griss J, Martín M, O'Donovan C, *et al.*: **Consequences of the discontinuation of the International Protein Index (IPI) database and its substitution by the UniProtKB "complete proteome" sets.** *Proteomics.* 2011; **11**(22): 4434–4438.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Perez-Riverol Y, Vizcaíno JA: **Synthetic human proteomes for accelerating protein research.** *Nat Methods.* 2017; **14**(3): 240–242.
PubMed Abstract | Publisher Full Text

33. Southan C: **Supplementary data for assessing the human canonical protein count.** *figshare.* 2017.
Data Source

Open Peer Review

Current Referee Status: ✔ ? ?

Version 1

Referee Report 09 May 2017

doi:10.5256/f1000research.11995.r22248

? **Sylvain Poux** , **Lionel Brueza**
Swiss Institute of Bioinformatics, Lausanne, Switzerland

The article compares the number of canonical proteins encoded by the human genome in different resources, including UniProtKB/Swiss-Prot, HGNC, neXtProt, GeneID, Ensembl or CCDS. The major conclusion is that the number of canonical proteins should be around 19,000 and that, while numbers converge across resources the full canonical human proteome is still not defined.

This is a good description of the current situation and the article is therefore interesting even if there could be confusion between protein-coding genes and canonical proteins. If the author assumes it is the same, maybe for consistency reason it would be good to mention only protein-coding genes or to explain what the differences are. The author also suggests that an inter-team collaboration could come-up with a finished canonical proteome and seems to ignore the ways the different resources already collaborate. As this has already been raised in the review of E.Bruford, we will not enter into details. The question of the release cycle is also important and should be developed in more details. Many discrepancies are only transitional and only due to the release cycle of the resources compared. As mentioned in the article, neXtProt is built on UniProtKB/Swiss-Prot and differences between these resources are only due to release schedule. But this is also holds true for the other resources and should be emphasized.

Another issue concerns the methodology of the study. A number of resources compared in this study do not have the same primary mission and it is therefore normal to have discrepancies between them. For example, HGNC is a nomenclature committee and official gene names are assigned when a consensus name is reached in the community. As a consequence, some clear protein-coding genes, such as NSG1 and NSG2 (UniProt P42857 and Q9Y328, respectively) are not yet present in HGNC, because no consensus has been found for these genes. The same is true for CCDS, which aims to provide a consensus sequence for all protein-coding genes: some protein-coding genes are absent from the CCDS set because no consensus has been found for the sequence (for example ELOA3C; UniProt A0A087WX78).

An alternative approach to assess the number of human protein-coding genes might be to compare portals described in this article with proteomics resources: it might be interesting to investigate the number of peptides that do not match to protein-coding genes in HGNC, UniProtKB/Swiss-Prot or GeneID.
We think that the article would benefit developing these different points in the discussion.

There are a number of typos and imprecisions in the text listed below that alter the quality of the manuscript and should be reviewed:

“are all cross referenced to a single, maximal length, protein entry.”
It is not absolutely true since the maximal length is one the criteria. However, the relevance of the selected canonical protein in Swiss-Prot in term of expression and biological relevance are also considered among other criteria.

“the fact that that”

“the first two are essentially automated pipelines” it is not clear what the author is referring to? Swiss-Prot and HGNC?

“There is now a community effort to promote more proteins to P1”
The author uses indifferently PE1 to PE5 and P1 to P5. This could be misleading.

“indicate the correct human sequence
is the 35 resides represented in”

One should read residues instead of resides.

“The current UniProt has”
When mentioning the database, prefer UniProtKB

Is the topic of the review discussed comprehensively in the context of the current literature?
Yes

Are all factual statements correct and adequately supported by citations?
Partly

Is the review written in accessible language?
Yes

Are the conclusions drawn appropriate in the context of the current research literature?
Partly

Competing Interests: We are working for UniProtKB/Swiss-Prot

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Referee Report 05 May 2017

doi:10.5256/f1000research.11995.r21692

? **Elspeth Bruford**
European Bioinformatics Institute, European Molecular Biology Laboratory, Hinxton, UK

General remarks:

The author has investigated the question of how many protein coding genes are encoded in the human genome, and come to the conclusion that while protein coding gene counts from a variety of resources do appear to be converging, there are still significant differences. One key aspect the author has maybe not fully appreciated is the considerable level of collaboration already occurring between the cited resources, which can be both advantageous - by reaffirming correct annotations - and disadvantageous - by perpetuating erroneous annotations through multiple resources. At the same time, definitions of biotypes, and membership within in each biotype, do still present differences which groups, including members of the CCDS collaboration, are looking to unify. Furthermore, while different interpretations of available data can of course cause discrepancies (and this is perhaps an area where more work is required by the community to reach agreed standards, for example see PMID 26367542[1]), unsurprisingly some resources access different datasets which cause further differences. While collaborations, definitions and data-sharing could be tightened up, there is no doubt that what is most needed is concerted experimental investigation of the remaining putative/hypothetical/dubious protein coding loci that remain within the genome, so that the resulting data can be used to decide upon a definitive biotype for these loci.

Overall this is discussing an important question but there is a tendency to be rambling in sections and I think the paper needs better organising to highlight some interesting questions the author raises. Some assumptions made about the various projects also need to be corrected, and more attention to detail is required for the numbers quoted to avoid confusing readers.

Specific remarks:

use of "miss-" throughout instead of mis-
Ensembl, not Ensemble
Entrez Gene is more widely referred to now as "NCBI Gene" - but variously referenced throughout the ms as "GeneID", "NCBI genome annotation", "RefSeq and Gene", "NCBI Entrez Gene", "GI", "NCBI", "NCBI pipeline automation" etc...
GENCODE, not GENECODE
HGNC is HUGO Gene Nomenclature Committee (not Human)
neXtProt, not neXtprot

Abstract:

I disagree that the only suggestion that total numbers of protein coding genes may rise is from reports of smORFS - as the author discusses later in the paper, even the very few genes reported to date to encode "smORFs" have limited evidence. I would anticipate most increase would come from careful re-annotation using increasing amounts of data (conservation, RNAseq, etc) and from annotation of multiple haplotypes that may cover regions of the human genome that are not currently included in the reference assembly and could be included in the future as alternate loci by the GRC.

Introduction:

While saying that the longest mRNA strategy has data support, it would also be worth mentioning the exception of read-through transcripts which can confuse this strategy significantly.

Historical Growth:

It could also be worth noting that Ensembl and Swiss-Prot/UniProtKb are also coupled as Ensembl sequences that are absent from UniProtKB are imported into UniProtKB/TrEMBL and tagged as part of their human proteome.

Ensembl's statistics make it very clear the number of proteins encoded by readthrough transcripts and on "alternative sequence", so I don't see how these could be said to "complicate" the figures. The issue of how many of the proteins (and protein coding genes) included on the alt loci are not represented in the primary assembly is however an interesting question.

Figure 2 shows the Swiss-Prot protein counts divided into total in red and those with protein or transcript evidence in blue - it would be nice to have the 2017 figures actually stated as opposed to having to guesstimate them from the graph.

Current Counts:

The author does not seem to understand the relationship between GENCODE (not GENECODE), Vega/Havana and Ensembl. It is nicely explained on the GENCODE site: https://www.gencodegenes.org/faq.html

Hence none of these figures are truly independent at all, and any differences between Ensembl and GENCODE figures are likely due to release asynchrony.

As it is unsurprising that GeneCards, which combines data from a variety of resources, has the largest "protein coding gene count" it is equally unsurprising that the CCDS consortium has the lowest as they are looking for the consensus CDS from Ensembl/Havana (=GENCODE) and RefSeq.

I disagree with the statement that mapping identifiers across sources can "establish if the protein sequence in pipeline output A is the same as pipeline B", and indeed the author discusses the example of *BACE1* which shows this is not necessarily true; however, this discrepancy is not due to the mappings themselves or how they are made, but simply due to the methods of protein prediction/selection used in each resource. The mapping may correctly suggest that both pipelines are considering the same genomic locus (in this case the *BACE1* gene), but agreement on the encoded protein(s) is not guaranteed. This paragraph would be better rephrased to make this clear.

Typo: if HGNC instantiate then they should also collate (not collates)

Cross-reference counting:

"However, the choice was made here to exemplify just four identifiers, Swiss-Prot accession numbers, HGNC IDs (directly, or via the current gene symbols) Ensembl gene IDs and NCBI Entrez Gene IDs. These were chosen for their global prominence but also methodological complementarity. This derives from the fact that that the first two are essentially automated pipelines (but different), while the second two are primarily manual expert annotation operations (but also different)"

The first two resources in the list are Swiss-Prot and HGNC, and neither are "essentially automated"; I think the author meant to say "last two" as nowhere else in this paper does he suggest either of these resources rely heavily on automation. As discussed earlier, the Ensembl gene set is a merge of their automated predictions with Havana manual annotations, and the manual annotations make up the vast majority of the protein coding genes. Likewise, the NCBI "Entrez Gene IDs" undergo extensive manual

curation, especially for the human set. Therefore I would disagree that *any* of these four resources are "essentially automated pipelines".

Figure 3 is very confusing with all of the "zero" segments - this figure would make more sense if it was an "all by all" comparison, as opposed to being based solely on the Swiss-Prot dataset. In fact I would venture that a simple table would be more readable, and as stated later in the paper this is a "Venn-type set(s) that generally end(s) up being more confusing than illuminating." Further, the numbers listed in Fig 3 do not correspond with those in the text - in Table 1 20,617 mappings were listed for "GeneID"/NCBI, whereas here there are 18,896, a difference of 1,721, not 2,923 as stated in the text. And for HGNC 19,957 is 924 higher than the 19,033 listed in Table 1, not 905 higher as stated.

The reason for the increase in mappings to HGNC IDs is more likely the inclusion of mappings to loci that HGNC do not regard as protein-coding, such as immunoglobulin light chain segments, than it is due to Swiss-Prot having more than one HGNC ID in any given record. In fact this explanation is discussed in the next paragraph where the types of loci enriched in specific segments are discussed, such as immunoglobulin light chain segments and endogenous retroviruses (note again the NCBI pipeline is referred to as "automation" which I do not think is a fair representation). I think it would make more sense for these two paragraphs to be rewritten to present the reasons more coherently. Perhaps it also would have made a better comparison to limit the Swiss-Prot data to loci that ALL four resources regard as protein coding, or to at least present how many of the loci in some segments of the diagram each resource individually considers as protein coding? This would also have made a comparison with the figures in Table 1 more valid, as currently the figures in Table 1 represent a different set to those being compared in Figure 3. Finally, the zero figures in the Ensembl (not Ensemble!)/Swiss-Prot and NCBI/Swiss-Prot segments must reflect their efforts to map between resources, though I am surprised there are no differences at all, even due to update cycle asynchrony? These figures certainly do not result from HGNC importing everything that NCBI and Ensembl annotate automatically (which I note Michael has suggested in his review), as this is definitely not the case. In the next paragraph 19,035 rows are quoted (twice) for HGNC data, but again this does not tally with the figure of 19,033 quoted in Table 1 for HGNC protein coding loci.

Existence Evidence:

In the figures quoted with evidence from Peptide Atlas for Swiss-Prot and neXtProt (17,084 vs 18,083) I would disagree that this could be described as a "slight" difference; this is nearly 1000 loci, which is at least 5% of the protein coding loci in the genome, even using the highest of the counts cited in this paper. What is the reason for this difference, it would be interesting to know. In the next paragraph the figure of 152 is quoted for the HPA-only set, but from Figure 5 this looks to be 158. Which is correct?

Typo - "...complications include the 40-residue **of** putative protein FAM86JP...".

Also note that while FAM68JP does not have cross-reference to NCBI Gene or Ensembl from Swiss-Prot these can be found in HGNC and NCBI Gene. The last paragraph of this section again mentions the issue of IG chains, which most resources do not class as "protein coding".

Small Proteins:

HGNC symbol is APELA (not APLEA).
Typo: "...to generate P0DN83 and P)DN84 for **a** 34 residue mouse and human proteins..."

The name of the second smORF (DWORF) is not actually mentioned until it is listed in the bullet points, it would be good to introduce "DWORF" by name in the paragraph above. I do not agree with the author that from the (paucity of) evidence cited for these examples that it is therefore "certain" that additional smORfs will be discovered, I think "likely" would be more appropriate.

Data Availability:

When discussing data update cycles, note that HGNC have daily updates and I think the same can be said of NCBI Gene, so these are both far more frequently than monthly.

Again the figure quoted in the text does not match numbers given in Figure 3: this section says that the query for HGNC cross-references gave 19967, while Fig. 3 quotes 19957.

Typo: "...pre-selects are not necessary since **is** human Swiss-Prot derived..."

References

1. Bruford EA, Lane L, Harrow J: Devising a Consensus Framework for Validation of Novel Human Coding Loci.*J Proteome Res*. 2015; **14** (12): 4945-8 PubMed Abstract | Publisher Full Text

Is the topic of the review discussed comprehensively in the context of the current literature?
Yes

Are all factual statements correct and adequately supported by citations?
Partly

Is the review written in accessible language?
Yes

Are the conclusions drawn appropriate in the context of the current research literature?
Partly

Competing Interests: I am the Project Coordinator and one of the PIs of the HGNC project

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Referee Report 18 April 2017

doi:10.5256/f1000research.11995.r21691

Michael Tress
Structural Biology and Bioinformatics Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

Abstract:
“In 2004, when the protein estimate from the finished human genome was only 24,000, the surprise was compounded as reviewed estimates fell to 19,000 by 2014. “
This makes no sense; it seems to be missing a large chunk.

“miss-annotation“

Introduction:
“This has its origins in the Swiss-Prot approach to protein annotation whereby protein sequence differences arising from the same genomic locus either by alternative splicing or alternative initiations (or permutations of both) and/or genetic variants, are all cross referenced to a single, maximal length, protein entry[8].“

This is not strictly true, SwissProt does NOT divide up all proteins from the same gene in different entries (TMPO for example). Here you have to be clear that SwissProt does this most of the time.

“Importantly, while this was originally introduced as the curatorial strategy of choosing the longest mRNA for an entry, it actually turns out to have post- genomic data support, not only in the form that coding-loci express a single main protein (i.e. that most predicted alternative transcripts may not be translated), but also that in most cases this is the max-exon form (i.e. the curatorial choice actually seems to be the biological “default”)[9].”

Strictly speaking this is true, the longest SwissProt form is the biological default in most cases. But it is purely technical and is not the best way of selecting the biological default. The way this paragraph is written makes it sound like it is. Better to say:

“not only in the form that coding-loci express a single main protein [ref to Ezkurdia et al, JPR] (i.e. that most predicted alternative transcripts may not be translated), but also that in most cases this max-exon form (i.e. the curatorial choice) actually coincides with the biological “default”)[9].”

Ref: Ezkurdia I, Rodriguez JM, Carrillo-de Santa Pau E, Vázquez J, Valencia A, Tress ML. Most highly expressed protein-coding genes have a single dominant isoform. J Proteome Res. 2015 Apr 3;14(4):1880-7. doi: 10.1021/pr501286b. [1]

Historical Growth:
“One of these is the inclusion of “alternative sequence”, referring to genomic sections that differ from the primary contiguous assembly. The current release of Ensembl (87.38) species 2,541 proteins in this category, but it is not clear which of these are just variants of those derived from the primary assembly.”

Alternative sequence genes are not included in the Ensembl reference counts.

principle = principal

GENCODE not GENECODE!

And GENCODE, VEGA and Ensembl ARE merged and have been for a number of years.

VEGA is annotated by the HAVANA group (part of the GENCODE Consortium), not Havanna.

It's also worth pointing out that Ensembl (since it is now merged with GENCODE) is essentially a manually curated annotation too with manual curations coming from the HAVANA team.

"Consensus Coding Sequence (CCDS) project. These correspond to a core set of proteins annotated as having full length transcripts that exactly match reference genome coordinates."

In fact CCDS transcript models need to exactly match between RefSeq and Ensembl/GENCODE, which explains why CCDS is the smallest set. This is actually an important caveat for the next paragraph, as might be imagined.

Cross-reference counting:
It is also worth mentioning that these are the only four independent sets, in that Vega and GENCODE merge into Ensembl, NextProt is UniProt and GeneCards and CCDS are essentially intersections and unions of different subsets.

Ensembl, not Ensemble.

"The explanation is that Ensemble and NCBI Gene have consolidated mapping reciprocity for proteins in Swiss-Prot (but, as mentioned above, many proteins from these two sources are still nominally "outside" Swiss-Prot)."

I think what it really says is that genes annotated in both Ensembl and SwissProt are automatically included in HGNC.

"GENCODE and Vega do not currently have cross-references inside Swiss-Prot"

Because GENCODE/VEGA == Ensembl

"forth" - fourth

Existence evidence:
"However, on its own, active transcription is insuficient to prove translation, even with a predicted CDS"

Maybe given the proliferation of such papers it might be worth pointing out that neither is ribosome profiling evidence …

"in regarded to" in regard to

"As was done for Figure 3, "
I think this whole paragraph could be written more carefully. I can follow it, but I suspect most people wouldn't. The data sets being compared need to be introduced specifically (again) and the numbers cross-checked. The examples are interesting, but:

"A second example exposes a different problem. The putative uncharacterized protein C7orf76 (Q6ZVN7) is mapped from UniProt to a different protein in HPA as ENSG00000127922- SHFM1 (i.e. P60896). The

miss-mapping appears to be extrinsic to HPA and in this case could be a UniProt < > Ensembl problem (which is why this is not in the 4-way set)."

Actually the problem stems from the fact that Ensembl annotates a single gene (now called SEM1) for these coordinates, while RefSeq has two (those listed in the paper). I have looked at this case before and wrote "RefSeq has two genes for SHFM1; RefSeq is right". I am not 100% sure that it is, but if it is one gene, it looks to be a gene that has two ORFs and hence it makes sense that UniProt has two entries.

Small proteins
"APLEA" = APELA

Also worth pointing out the conservation all the way back to Danio for APELA. In fact cross-species conservation studies currently being undertaken by Ensembl may unearth some "missing" smORFs.

"not withstanding miss-matches"

"miss-mapping"

"However, inter-source clustering of explicit protein sequences could make identifying difference more effectively than cross- references alone (e.g. a possible resurrection of the Human Protein Index initiative [31])."

Nooooo, do not resurrect the IPI, it died for good reasons. I don't believe that we need any more competing (and only superficially communicating) bodies in the field.

"It could be argued that additional (collective) manual curation would be needed to accomplish this"

This is nice in principal, but manual curation is VERY subjective. For many genes whether it is annotated as coding or not is based on the balance of probabilities and each annotator has his/her own balance of probabilities. What is needed is more and better information for the corner cases.

General Comment:
There are a lot of distinct sets being compared in the figures. I would name and define the sets clearly in the text when possible, otherwise readers will struggle to see what is being compared.

References
1. Ezkurdia I, Rodriguez JM, Carrillo-de Santa Pau E, Vázquez J, Valencia A, Tress ML: Most highly expressed protein-coding genes have a single dominant isoform.*J Proteome Res*. 2015; **14** (4): 1880-7
PubMed Abstract | Publisher Full Text

Is the topic of the review discussed comprehensively in the context of the current literature?
Yes

Are all factual statements correct and adequately supported by citations?
Partly

Is the review written in accessible language?
Yes

Are the conclusions drawn appropriate in the context of the current research literature?
Yes

Competing Interests: I am part of the GENCODE Consortium

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.