HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics

Rui M M Branca[1,6], Lukas M Orre[1,6], Henrik J Johansson[1,6], Viktor Granholm[2], Mikael Huss[3], Åsa Pérez-Bercoff[1], Jenny Forshed[1], Lukas Käll[4,5] & Janne Lehtiö[1]

We present a liquid chromatography–mass spectrometry (LC-MS)-based method permitting unbiased (gene prediction–independent) genome-wide discovery of protein-coding loci in higher eukaryotes. Using high-resolution isoelectric focusing (HiRIEF) at the peptide level in the 3.7–5.0 pH range and accurate peptide isoelectric point (pI) prediction, we probed the six-reading-frame translation of the human and mouse genomes and identified 98 and 52 previously undiscovered protein-coding loci, respectively. The method also enabled deep proteome coverage, identifying 13,078 human and 10,637 mouse proteins.

Current approaches to genome annotation begin with analysis at the DNA level, using noncomparative (*ab initio* gene prediction) and comparative (sequence similarity across species) methods to identify candidate gene models; these are followed by validation based on expressed RNA sequences. Yet the definite proof of a genomic locus being protein coding is the detection of its corresponding protein. Hence, it is desirable to integrate large-scale proteomics data into gene annotation, as is done in the field of proteogenomics[1]. This endeavor requires the construction of protein databases from the six-reading-frame translation (6FT) of the genome for the purpose of matching spectra from tandem mass spectrometry (MS/MS). Proteogenomics has already proven its value in organisms with small genome size, such as prokaryotes[2]. In higher eukaryotes, the large genome size and low protein-coding content result in decreased sensitivity (number of identifications) and specificity (number of correct identifications)[1,3]. As protein sequence databases (the search space) expand, the scores for the highest-scoring incorrect spectrum matches inevitably increase, whereas the score of correct spectrum matches remains invariant[4]. In practice, this severely lowers the sensitivity of database searching against 6FTs of large genomes.

To our knowledge, a full 6FT search of a higher eukaryotic genome has been performed on only *Arabidopsis thaliana* (genome size of 157 Mb)[5]. In this case it was possible to tackle the entire 6FT search space, which is infeasible for species with larger genomes, such as *Homo sapiens* (genome size of 3.2 Gb and ~1.5% protein-coding content) or *Mus musculus* (genome size of 3 Gb) (**Fig. 1a**). Thus, proteogenomics efforts on organisms with large genomes require extensive search-space reductions. This has been attempted through the use of gene prediction algorithms, RNA-seq or expressed sequence tag databases, and via database fractionation[3,6–11].

Here we combine experimental prefractionation based on peptide pI and theoretical prediction of pI to achieve search-space reduction. For peptide prefractionation we developed HiRIEF, in this work targeting the 3.7–5.0 pH range of the entire tryptic peptide population, which represents about one-third of the human tryptic peptidome (according to the theoretical pI distribution) (**Fig. 1b**). Lysates prepared from human A431 cells and mouse N2A cells (**Supplementary Fig. 1**) were digested with trypsin, and the resulting peptide samples were prefractionated using HiRIEF gel strips (narrow range, pH interval 3.7–4.9; ultranarrow range, pH intervals 3.70–4.05, 4.00–4.25, 4.20–4.45 and 4.39–4.99), each divided into 72 fractions (Online Methods). Peptide-level isoelectric focusing (IEF) has previously been used in mass spectrometry as a prefractionation method with varying degrees of resolution[12–15]. Using ultranarrow HiRIEF strips, we achieved at least a fivefold increase in resolution (0.0035 pH units per fraction) compared to previously reported values. Each fraction was analyzed by LC-MS on an LTQ Orbitrap Velos (Thermo Scientific).

We initially evaluated the HiRIEF LC-MS performance in a conventional search against the reference proteome, identifying 13,078 and 10,637 proteins in human and mouse, respectively (**Supplementary Table 1**). Notably, of the 77,785 distinct human peptides identified, 39,941 were previously not present in the human subset of Peptide Atlas (http://www.peptideatlas.org/, release 2012-04) corresponding to 18% of all the peptides in that subset of the repository (**Supplementary Fig. 2**). We also investigated the identification gain of using subcellular fractionation coupled with HiRIEF fractionation (**Supplementary Figs. 3** and **4**). HiRIEF was shown to be robust in terms of peptide fractionation resolution, complexity reduction and reproducibility (**Supplementary Figs. 5**–**7**). The redundancy between HiRIEF fractions was minimal, with the majority of peptides being well resolved and identified in one or at most two consecutive fractions (**Supplementary Fig. 5**). The capability of HiRIEF to reduce

[1]Department of Oncology-Pathology, Science for Life Laboratory, Karolinska Institutet, Stockholm, Sweden. [2]Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden. [3]Department of Biochemistry and Biophysics, The Arrhenius Laboratories for Natural Sciences, Science for Life Laboratory, Stockholm University, Stockholm, Sweden. [4]School of Biotechnology, Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden. [5]Swedish e-Science Resource Center, KTH Royal Institute of Technology, Stockholm, Sweden. [6]These authors contributed equally to this work. Correspondence should be addressed to J.L. (janne.lehtio@ki.se).

RECEIVED 22 MARCH; ACCEPTED 8 OCTOBER; PUBLISHED ONLINE 17 NOVEMBER 2013; DOI:10.1038/NMETH.2732

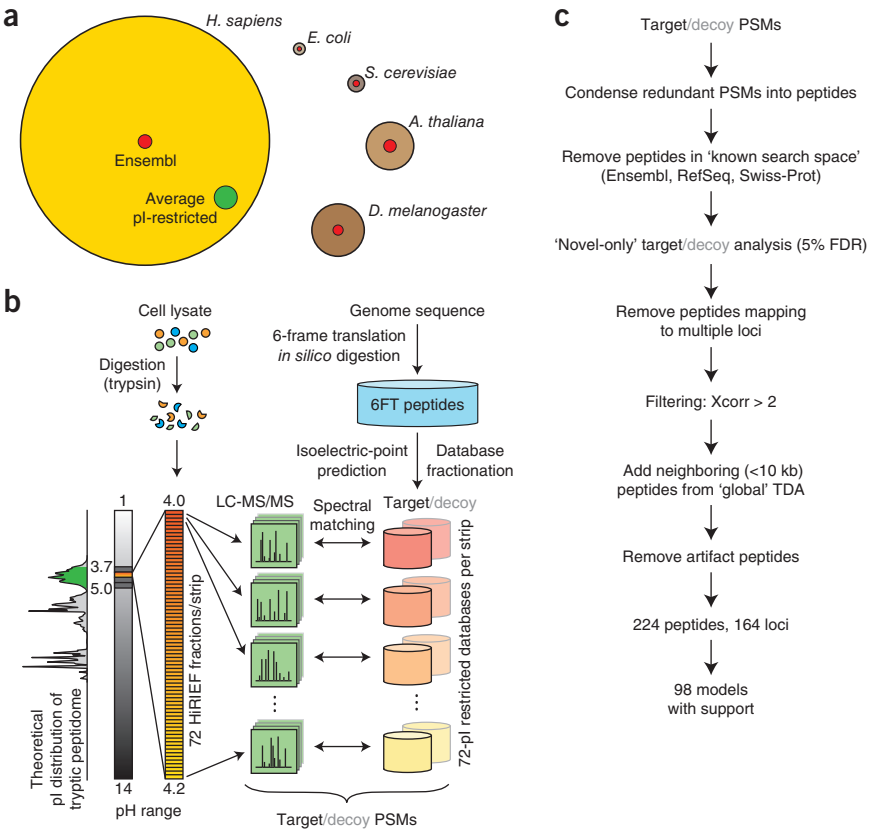

Figure 1 | HiRIEF LC-MS enables unbiased proteogenomics in higher eukaryotes. (**a**) Search spaces resulting from the 6FT and theoretical tryptic digestion of genomes from different organisms (circle areas are proportional to the number of tryptic peptides). The red core of each circle represents the peptide search space arising from the reference proteome for each organism (all from Ensembl, except for *A. thaliana*, which is from The Arabidopsis Information Resource, http://www.arabidopsis.org/). The average size of pI-restricted databases generated by our workflow is indicated in green. (**b**) Overview of the proteogenomics workflow. Extracted proteins are digested, and the resulting peptides are fractionated using different IEF strips (each producing 72 fractions) covering the pH range between 3.7 and 5.0. Following LC-MS/MS analysis, spectral matching is done between each fraction and its corresponding pI-restricted target/decoy database. The pI-restricted databases were generated with the aid of the pI prediction algorithm PredpI (**Supplementary Software**). (**c**) Strategy for discovery of novel peptides. The displayed numbers pertain to the proteogenomics experiment in *H. sapiens* (**Supplementary Fig. 11**). The same workflow was used for *M. musculus*. TDA, target-decoy analysis.

sample complexity and increase information content is demonstrated by the even distribution of peptides over the 72 fractions of each strip (**Supplementary Fig. 6**). The high reproducibility of IEF is shown by comparing the focusing position of peptides between different runs (**Supplementary Fig. 7**). The analytical depth of the proteomics experiment in the human cell line was comparable to that of transcriptomics data we generated by RNA-seq on aliquots from the same samples (372 million paired-end reads) (**Supplementary Fig. 8**).

For proteogenomics analysis, we developed an approach based on peptide pI prediction in order to search the MS raw files against the full human genome in an unbiased fashion (**Fig. 1**). First, *in silico* tryptic digestion was performed on the full 6FT of the human genome, resulting in 257 million tryptic peptides (splice junction–spanning peptides were not included). Because the matching of MS raw data against this search space is unwieldy, we used peptide pI information generated by HiRIEF fractionation to achieve rational reduction of the search space. We divided the full 6FT database of tryptic peptides into 360 pI-restricted databases corresponding to the fractions generated by HiRIEF (5 pH ranges × 72 fractions) using an in-house–developed algorithm for peptide pI prediction (PredpI; **Supplementary Software**). The pI interval of each pI-restricted database was centered at the middle pI value of the corresponding HiRIEF fraction. The interval width was determined by the experimental fraction pI width plus the prediction error margin (±0.06), which was based on the prediction accuracy evaluation (**Supplementary Fig. 9**).

The full search space was divided into the categories 'known search space' (i.e., for peptides contained in Ensembl, RefSeq or Swiss-Prot human protein databases; in total 0.3% of the entire search space) and 'novel search space' (the remaining 99.7%). In the 'global' target-decoy analysis (TDA), the vast majority of peptide spectrum matches (PSMs) corresponded to the known search space (the known peptides identified by this approach were compared to the conventional search results; **Supplementary Fig. 10**). Only 2.2% of PSMs were assigned to the novel search space. The false positives of the global TDA (with 1% false discovery rate (FDR), PSM level) are likely to be enriched in the novel search space because it is much larger than the known search space. One strategy that has been employed previously[5] to increase confidence in the novel protein-coding loci is to require at least two neighboring peptides. We used a different approach for FDR estimation of the peptides in the novel search space (**Fig. 1c** and **Supplementary Fig. 11**): all hits on the target database matching to the known search space were removed before a novel-only TDA (5% FDR, peptide level). This strategy made it possible to explore single novel peptide hits, which are important because in many instances there is room for only one novel tryptic peptide (for example, exons extended by only a few amino acids, short N-terminal extensions and very short proteins). An additional set of requirements was applied (9–30 amino acids, PSM score (Xcorr) > 2.0, and mapping to a single genomic locus), yielding 229 unique novel peptides. The novel genomic loci were further populated by the addition of neighboring peptides (<10 kb apart) from the global TDA, totaling 252 novel peptides. Twenty eight of these peptides could have originated from known proteins via unusual phenomena (i.e., cleavage before proline residues, artifact deamidations and N-terminal shuffles) and were therefore discarded.

The distribution of precursor mass errors of novel peptide PSMs was compared and shown to be similar to that of known peptides (**Supplementary Fig. 12**). The MS2 spectra of novel peptides

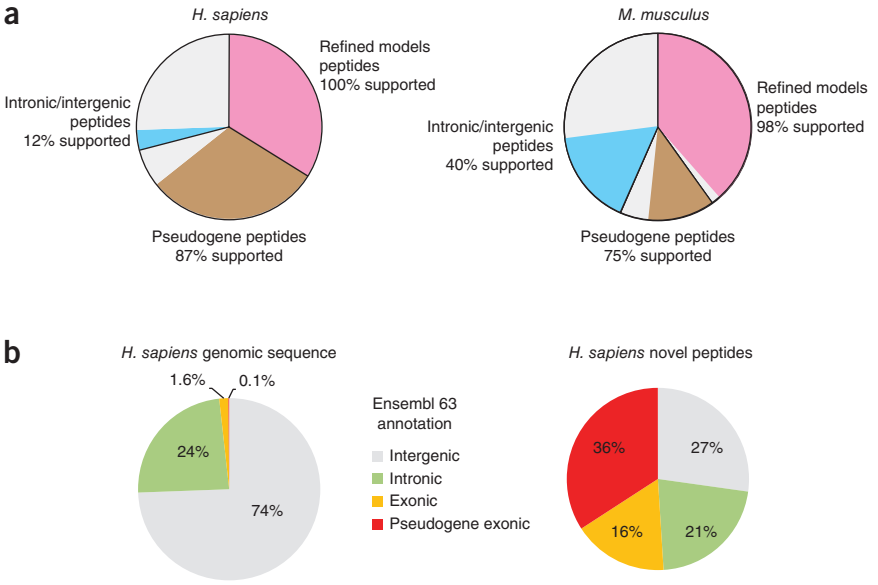

Figure 2 | Analysis of distribution of novel peptides into different categories. (**a**) Distribution of novel peptides into gene-model categories in *H. sapiens* (224 novel peptides; **Supplementary Table 2**) and *M. musculus* (122 novel peptides; **Supplementary Table 3**). A peptide is considered supported (colored sections) if it is covered in full extent by RNA-seq or a gene prediction algorithm. (**b**) Overrepresentation analysis: comparison of annotation category distribution for the human genomic sequence and novel peptides. For decomposition of the genomic sequence into intergenic, intronic, exonic and pseudogene exonic categories, nucleotides were counted in forward plus reverse directions using the Ensembl genome annotations (release 63).

were annotated and manually inspected (**Supplementary Data 1**). The 224 novel peptides were evaluated in terms of supporting evidence (**Supplementary Table 2**). Finally, to demonstrate the method's applicability to samples from other organisms, we applied the same proteogenomics workflow to MS data obtained from a whole-cell lysate of a mouse cell line (N2A) and obtained 122 novel peptides (**Supplementary Table 3**).

The 224 human and 122 mouse novel peptides cluster into 164 and 101 genomic loci, respectively. For evaluation of the novel protein-coding loci, several different types of independent evidence were used, including RNA-seq, mammalian conservation, gene prediction and data from previous studies. The genomic loci can be grouped into different classes, discussed below: (1) refined models of known genes (47 human, 32 mouse), (2) pseudogenes and long noncoding RNA genes (51 human, 20 mouse) and (3) intronic and intergenic loci with no connection to gene annotations (66 human, 49 mouse) (**Fig. 2** and **Supplementary Tables 2** and **3**).

1. The suggested refined models were built using peptide and RNA-seq data and were further categorized into exon bridging, new exons or exon extensions, N-terminal extensions, alternative stops, or genes containing alternative reading frames or upstream open reading frames (uORFs) from the canonical coding DNA sequence. Moreover, there was evidence for near-cognate translation initiation in many instances. Interestingly, ribosomal profiling has recently confirmed widespread translation of uORFs, which often use near-cognate initiation sites[16]. Examples of refined gene models are shown in **Figure 3** and **Supplementary Figures 13** and **14**.

2. Pseudogenes represent less than 0.1% of the total search space, yet a surprisingly large number, 36%, of human novel peptides mapped to pseudogenes (**Fig. 2b**). These findings are supported by recent peptide-level evidence of pseudogenes in mouse[6]. In humans, the observation of lineage- and cancer-specific expression of pseudogenes at the RNA level indicates biological relevance[17]. Our data suggest

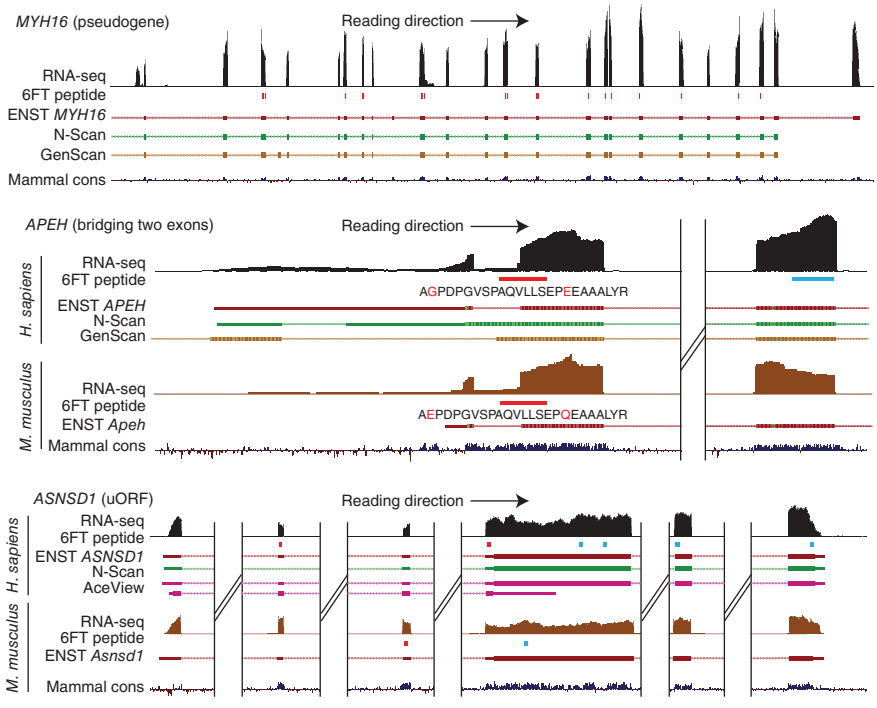

Figure 3 | New gene models. Examples of new gene models in *H. sapiens* (*MYH16*) and new gene models found in orthologous genes of *H. sapiens* and *M. musculus* (*APEH*, a bridging of two exons, and *ASNSD1*, an upstream open reading frame (uORF) containing multiple exons). The sequence of novel peptides found is displayed for comparison between species where relevant. RNA-seq reads (human) and peptides identified in the proteogenomics experiment ('6FT peptide'; novel peptides are shown in red and known peptides in blue) as well as public-domain data—RNA-seq reads (mouse), Ensembl annotations ('ENST'), gene predictions (N-Scan, GenScan and AceView) and mammal conservation scores (PhyloP; 'Mammal cons')—are shown. The images were generated using the UCSC Genome Browser (http://genome.ucsc.edu/) with the human February 2009 (GRCh37/hg19) assembly and the mouse July 2007 (NCBI37/mm9) assembly.

that pseudogenes may be not only transcribed but also translated. An interesting particular example was the pseudogene *MYH16*, identified by 20 peptides (**Fig. 3**), which were validated by LC-MS using synthetic peptides (**Supplementary Fig. 15**). The protein-coding capacity of *MYH16* was previously shown to have been lost through double base deletion (resulting in a premature stop codon) during divergence of the human lineage from other primates[18]. However, our data show that, in the A431 cell line, the *MYH16* gene is actively encoding a shorter protein isoform with its translation initiation site downstream from the aforementioned double base deletion.

3. For unconnected intronic and intergenic loci, our RNA-level data did not link the found peptides to Ensembl gene annotations. This class showed an overall low level of support from independent data. Considering also the distribution of search space (**Fig. 2b**), this category most likely contains the majority of false positive hits in the proteogenomics search.

Strikingly, when comparing the results of our human and mouse proteogenomics analyses, we discovered five identical new gene models in orthologous genes (in *H1FX*, *APEH*, *CNOT1*, *ASNSD1* and *CRYBG3*), found independently in human and in mouse (**Fig. 3** and **Supplementary Fig. 14**). For *CRYBG3*, we found 16 peptides in human and 5 peptides in mouse, which indicated that this gene should absorb its upstream neighbor in both species. During the preparation of this manuscript, *CRYBG3* annotation was corrected for *H. sapiens* in RefSeq. These findings suggest that proteomics data can correct some systematic errors in genome annotation.

In summary, using the proteogenomics approach described here, we found evidence for 98 new protein-coding loci in *H. sapiens* (47 refined gene models, 50 pseudogenes and 1 long noncoding RNA gene) and 52 in *M. musculus* (32 refined gene models, 19 pseudogenes and 1 antisense gene).

Recent progress made by high-throughput RNA sequencing suggests that as much as 74.7% of the entire human genome is transcribed[19]. Moreover, recent work on human evolutionary constraints suggests the existence of as much as 2% novel protein-coding exons[20], indicating that the human genome annotation is not complete.

The benefit of employing peptide pI–based fractionation in ultranarrow ranges is twofold. First, it provides outstanding peptide resolving power, thereby reducing complexity and allowing deep proteome coverage. The depth of the method is demonstrated by broad coverage of the reference proteome in both human and mouse. Second, the pI for each peptide can be predicted with sufficiently high accuracy, allowing a rational fractionation of the databases used for peptide spectral matching, thereby permitting gene prediction–independent proteogenomics. In order to enable HiRIEF proteogenomics in the full pI range of the human peptidome, extensive HiRIEF data on peptides with pIs in the 5–10 pH range need to be obtained. The PredpI algorithm can then be trained in the neutral and alkaline pH ranges to ensure high prediction accuracy.

With an ever-increasing rate of genomes being sequenced, there is a demand for new tools in gene annotation to define coding regions and protein variants. HiRIEF LC-MS enables both comprehensive conventional proteomics and unbiased proteogenomics in higher eukaryotes. We believe the method will have broad applicability for annotating the protein-coding genome.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. ProteomeXchange: MS raw and processed data files, PXD000065.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

Funding from the Swedish Research Council, Swedish Cancer Society, Stockholm's county council, Stockholm's cancer society and EU FP7 project GlycoHit is gratefully acknowledged. Support by BILS (Bioinformatics Infrastructure for Life Sciences) and J. Boekel in publishing the MS raw files is gratefully acknowledged. We thank the SciLifeLab genomics facility for experimental support and J. Lundeberg for the A431 sequence data. We thank E. Bereczki (Karolinska Institutet) for the kind gift of the N2A cell line. We thank K. Lindblad-Toh for critical reading of the manuscript. We acknowledge the late B. Bjellqvist for his early contribution in the development of IPG-IEF and peptide pI prediction.

AUTHOR CONTRIBUTIONS

J.L., R.M.M.B., L.M.O., L.K. and H.J.J. conceived of and designed the experiments. R.M.M.B. and H.J.J. performed the IEF separations and MS analysis. M.H., L.M.O. and R.M.M.B. performed the data analysis of RNA-seq experiments. R.M.M.B. performed the peptide pI calculations. L.K., J.L., R.M.M.B. and V.G. designed the database restriction workflow and performed the 6FT searches. L.K. and V.G. designed the novel-only TDA approach. Å.P.-B. performed the single-nucleotide polymorphism data analysis and calculated Ensembl annotation statistics. R.M.M.B., L.M.O., H.J.J. and J.F. performed proteomics data analysis. R.M.M.B., L.M.O. and J.L. wrote the manuscript. All authors were involved in discussion of the manuscript and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Krug, K., Nahnsen, S. & Macek, B. *Mol. Biosyst.* **7**, 284–291 (2011).
2. Kelkar, D.S. *et al. Mol. Cell. Proteomics* **10**, M111.011627 (2011).
3. Fermin, D. *et al. Genome Biol.* **7**, R35 (2006).
4. Granholm, V. & Käll, L. *Proteomics* **11**, 1086–1093 (2011).
5. Baerenfaller, K. *et al. Science* **320**, 938–941 (2008).
6. Brosch, M. *et al. Genome Res.* **21**, 756–767 (2011).
7. Evans, V.C. *et al. Nat. Methods* **9**, 1207–1211 (2012).
8. Edwards, N.J. *Mol. Syst. Biol.* **3**, 102 (2007).
9. Bitton, D.A., Smith, D.L., Connolly, Y., Scutt, P.J. & Miller, C.J. *PLoS ONE* **5**, e8949 (2010).
10. Sevinsky, J.R. *et al. J. Proteome Res.* **7**, 80–88 (2008).
11. Tanner, S. *et al. Genome Res.* **17**, 231–239 (2007).
12. Beck, M. *et al. Mol. Syst. Biol.* **7**, 549 (2011).
13. Hörth, P., Miller, C.A., Preckel, T. & Wenz, C. *Mol. Cell. Proteomics* **5**, 1968–1974 (2006).
14. Lengqvist, J., Uhlen, K. & Lehtio, J. *Proteomics* **7**, 1746–1752 (2007).
15. Cargile, B.J., Sevinsky, J.R., Essader, A.S., Stephenson, J.L. Jr. & Bundy, J.L. *J. Biomol. Tech.* **16**, 181–189 (2005).
16. Ingolia, N.T., Lareau, L.F. & Weissman, J.S. *Cell* **147**, 789–802 (2011).
17. Kalyana-Sundaram, S. *et al. Cell* **149**, 1622–1634 (2012).
18. Stedman, H.H. *et al. Nature* **428**, 415–418 (2004).
19. Djebali, S. *et al. Nature* **489**, 101–108 (2012).
20. Lindblad-Toh, K. *et al. Nature* **478**, 476–482 (2011).

ONLINE METHODS

Protein extraction and subcellular fractionation. A431 human cells (DSMZ #ACC-91, tested for, and found free of, *Mycoplasma* contamination) were cultured in Dulbecco's Modified Eagle Medium supplemented with 10% fetal bovine serum, 1% penicillin/streptomycin, non-essential amino acids, and sodium pyruvate, all from Invitrogen. Twenty four hours after seeding, the A431 cell cultures (in triplicates) were treated with gefitinib to induce altered gene/protein expression. Controls were left untreated. At 2 h (treated and controls), 6 h and 24 h after treatment, cells were harvested by trypsination and washed in PBS. Cell suspensions were divided into aliquots for whole-cell analysis, subcellular fractionation and RNA sequencing. For whole-cell protein extraction, cells were lysed using buffer A (10 mM HEPES, pH 7.5, 10 mM KCl, 1.5 mM MgCl, HALT protease/phosphatase inhibitors, from Pierce, Thermo Scientific) supplemented with 2% SDS. After vortexing, samples were heated to 95 °C for 5 min and sonicated with a probe sonicator (Bandelin Sonopuls, Buch and Holm) twice using 50% duty cycle, 50% power for 15 s. For subcellular fractionation, cells were resuspended in buffer A and incubated on ice for 20 min. Cells were then disrupted by 30 strokes with a Kontes douncer (2 ml, pestle B small clearance 0.0005–0.0025 inch, Kimble Chase, Sigma-Aldrich). After centrifugation at 600*g* (5 min, 4 °C), the pellet (corresponding roughly to the plasma membrane and nuclear fraction) was washed once in buffer A, and proteins from the pellet were extracted by the method used for whole-cell extraction. The supernatant was centrifuged for pelleting of cytoplasmic organelles (100,000*g*, 1 h, 4 °C). The supernatant from the ultracentrifugation was supplemented with 1% SDS and saved as the soluble fraction. Proteins from the pellet were extracted by the method used for whole-cell extraction and saved as the organellar fraction.

N2A mouse cells (tested for, and found free of, *Mycoplasma* contamination) were grown on the same medium as above. After harvesting by trypsination, the N2A cells were lysed with a different buffer (4% SDS, 25 mM Tris, pH 7.5, 1 mM dithiothreitol), and proteins were extracted via acetone precipitation. Briefly, four volumes of ice-cold acetone were added to the cell lysate, which was then incubated at −20 °C for 1 h. After centrifugation at 12,000*g* and 4 °C for 10 min, the supernatant was discarded, and the pellet was washed once with 0.5 ml of ice-cold acetone. The pellet was redissolved in 1% SDS, which was followed by a stepwise dilution to 0.2% SDS and estimation of total protein concentration with the DC assay (Bio-Rad).

Protein digestion and iTRAQ labeling. A431 protein extracts from whole cell and from the three subcellular fractions were processed following a slightly modified FASP protocol[21]. The resulting peptide mixtures were arranged into several sets (see **Supplementary Fig. 1**) and labeled with iTRAQ8plex (AB Sciex). After pooling, all iTRAQ8-labeled peptide sets were cleaned by strong-cation exchange–solid-phase extraction (SCX-SPE, strata-X-C, Phenomenex) and divided into aliquots for IEF.

The N2A protein extract was reduced by 1 mM dithiothreitol, alkylated by 4 mM iodoacetamide, and digested with trypsin (Promega) overnight. The resulting peptide mixture was cleaned by SCX-SPE and divided into aliquots for IEF.

Peptide-level high-resolution isoelectric focusing (HiRIEF). The HiRIEF pH ranges employed were 3.7–4.9 (narrow range), 3.70–4.05, 4.00–4.25, 4.20–4.45, 4.40–4.65, and 4.39–4.99 (ultra-narrow range) (see **Supplementary Fig. 1**), all of them 24-cm-long IPG (immobilized pH gradient) gel-strip prototypes provided by GE Healthcare Bio-Sciences (GE Healthcare will provide these prototypes upon request). Each peptide mixture set (200 µg) was dissolved in 300 µl rehydration solution containing 8 M urea, which was used to re-swell the gel bridge overnight. In parallel, the IPG strip was re-swollen with 250 µl of 8 M urea, 1% IPG Pharmalyte, pH 2.5–5.0 (GE Healthcare). The gel bridge was applied at the cathode (acidic) end of the IPG strip, and IEF was run on an Ettan IPGphor (GE Healthcare) until at least 150 kV-h (for the 3.7–4.9 strips) or 250 kV-h (for the ultranarrow strips). After focusing was complete, a well-former with 72 wells was applied onto each strip, and liquid-handling robotics (GE Healthcare prototype) added MilliQ water and transferred the 72 fractions into a microtiter plate (96 wells, V-bottom, Corning cat. #3894), which was then dried in a SpeedVac.

LC-ESI-LTQ-Orbitrap MS/MS. Prior to each LC-MS run, the LC auto sampler (HPLC 1200 system, Agilent Technologies) dispensed 8 µl of solvent A to the dry HiRIEF fraction (in its microtiter plate well), mixed for 10 min, and injected 3 µl into a C18 guard desalting column (Zorbax 300SB-C18, 5 × 0.3 mm, 5-µm bead size, Agilent Technologies). We then used a 15-cm-long C18 PicoFrit column (100-µm internal diameter, 5-µm bead size, Nikkyo Technos) installed on to the nanoelectrospray ionization (NSI) source. Solvent A was 97% water, 3% acetonitrile (ACN), and 0.1% formic acid (FA); and solvent B was 5% water, 95% ACN, and 0.1% FA. At a constant flow of 0.4 µl/min, the curved gradient went from 2% B up to 40% B in 45 min, followed by a steep increase to 100% B in 5 min. Online LC-MS was performed using a hybrid LTQ Orbitrap Velos mass spectrometer (Thermo Scientific). Precursors were isolated with a 2-*m*/*z* window. We enabled "preview mode" for FTMS master scans, which proceeded at resolution of 30,000 (profile mode). Data-dependent MS/MS (centroid mode) followed in two stages: first, the top five ions from the master scan were selected for collision-induced dissociation (CID, at 35% energy) with detection in the ion trap (ITMS); and after, the same five ions underwent higher-energy collision dissociation (HCD, at 50% energy) with detection in the Orbitrap (FTMS). The entire duty cycle lasted ~3.5 s. Dynamic exclusion was used with 90-s duration. A precursor threshold of 1,000 counts was used. AGC targets were 1×10^6 for MS1, 2×10^4 for CID-MS2, and 5×10^4 for HCD-MS2. Injection-time maxima were 200 ms for CID and 500 ms for HCD.

Conventional proteomics using a reference proteome search. All MS/MS spectra were searched by Sequest/Percolator under the software platform Proteome Discoverer (PD, v.1.3.0.339, Thermo Scientific) using a target-decoy strategy. The reference databases used were the human protein subset of Ensembl 63 and the mouse protein subset of Ensembl 64. Precursor mass tolerance of 10 p.p.m. and product mass tolerances of 0.02 Da for HCD-FTMS and 0.8 Da for CID-ITMS were used. Additional settings were trypsin with one missed cleavage; Lys-Pro and Arg-Pro not considered as cleavage sites; carbamidomethylation on cysteine and

iTRAQ-8plex on lysine and N-terminus as fixed modifications; and oxidation of methionine and phosphorylation on serine, threonine or tyrosine as variable modifications. Quantitation of iTRAQ-8plex reporter ions was done using an integration window tolerance of 20 p.p.m. Peptides found at 1% FDR (false discovery rate) were used by the protein grouping algorithm in PD to infer protein identities. Target and decoy PSMs exported from PD (v.2.0, build229, Thermo) were processed through Mayu[22] to estimate protein-level FDR (1% cutoff was applied). The complete results of the conventional proteomics searches are reported in **Supplementary Tables 4** and **5**.

PredpI algorithm. The PredpI algorithm builds on a pI prediction algorithm used for proteins (explained in equations (1–4)) developed by Bjellqvist and coworkers[23] that is currently the basis of the Compute pI tool in Expasy (http://web.expasy.org/compute_pi/).

$$\mathrm{pH} = \mathrm{p}K + \log\left(\frac{[\mathrm{A}^-]}{[\mathrm{AH}]}\right) \quad \text{or} \quad \mathrm{pH} = \mathrm{p}K + \log\left(\frac{[\mathrm{B}]}{[\mathrm{BH}^+]}\right) \qquad (1)$$

Equation (1) is the Henderson-Hasselbalch equation for carboxylic acid groups (A, i.e., C termini and side chains of Asp and Glu) or amino groups (B, i.e. N termini and side chains of Lys, Arg and His).

$$\alpha_\mathrm{A} = \frac{[\mathrm{A}^-]}{[\mathrm{A}^-] + [\mathrm{AH}]} \quad \text{or} \quad \alpha_\mathrm{B} = \frac{[\mathrm{BH}^+]}{[\mathrm{BH}^+] + [\mathrm{B}]} \qquad (2)$$

α is defined as the charged fraction for each ionizable group (A or B) in a peptide.

$$\alpha_\mathrm{A} = \frac{-1}{10^{\mathrm{p}K-\mathrm{pH}} + 1} \quad \text{or} \quad \alpha_\mathrm{B} = \frac{-1}{10^{\mathrm{pH}-\mathrm{p}K} + 1} \qquad (3)$$

Equation (3) is derived from equations (1) and (2) and is used to calculate α for each ionizable group at a given pH.

$$\text{Net charge} = \sum_{i=1}^{\mathrm{A}} \frac{-1}{10^{\mathrm{p}K-\mathrm{pH}} + 1} + \sum_{i=1}^{\mathrm{B}} \frac{1}{10^{\mathrm{pH}-\mathrm{p}K} + 1} \qquad (4)$$

Equation (4) gives the net charge of a polypeptide at a given pH. The pI of a polypeptide is the pH at which the net charge of the polypeptide equals 0.

This theoretical framework (and the output from the Compute pI tool) is insufficient for accurate peptide pI prediction because it ignores the influence of neighboring charged groups on the p*K* value of each ionizable group. Cargile *et al.*[24] introduced the use of correction factors for the p*K* value of each ionizable group based on the influence of neighboring ionizable groups up to three residues away.

In PredpI, we further developed p*K* value correction by considering the influence of neighboring ionizable groups up to six residues away. In addition, each correction factor is multiplied by the charged fraction (α) of the neighboring ionizable group before being applied to the initial p*K* value. PredpI also introduces the use of a statistical correction factor that depends on the number and type (Asp or Glu) of carboxylic acid side chains existing in the peptide. Once the corrected p*K* values for all ionizable groups in a peptide are calculated using equation (5), they are then applied to equation (4) in order to obtain an accurately predicted peptide pI value.

$$\mathrm{p}K = \mathrm{p}K_0 + \sum_{i=1}^{I} \alpha_i \mathrm{dp}K_i + \mathrm{dp}K\mathrm{N} \qquad (5)$$

where p*K* is the corrected p*K* value for a given ionizable group, $\mathrm{p}K_0$ is the initial p*K* value, α_i is the charged fraction of the neighboring ionizable group *i*, $\mathrm{dp}K_i$ is the correction factor due to the influence of the neighboring ionizable group *i*, and dp*K*N is the statistical correction factor for carboxylic acid groups.

Original $\mathrm{p}K_0$ constants were obtained from Bjellqvist *et al.*[23] and refined using experimental HiRIEF data. dp*K* and dp*K*N correction factors were determined using experimental HiRIEF data. Optimization of p*K* constants ($\mathrm{p}K_0$, dp*K* and dp*K*N constants) was done by changing these constants in small discrete steps and verifying the impact on the prediction accuracy on a large training set. The final set of optimized p*K* constants was evaluated with a separate test set (**Supplementary Fig. 9**) and is provided in the **Supplementary Software**.

Discovery of new gene-model peptides using six-reading-frame translation. Nucleotide sequences of the 24 different human chromosomes were obtained from Ensembl GRCh37, release 63. The chromosomes were completely translated *in silico* using all six reading frames before being cut into peptides according to tryptic cleavage rules. Peptides of 6–40 amino acids in length, and with lysine or arginine residue on the C termini, were stored along with their chromosome positions. Furthermore, peptides with an acidic index (here defined as the number of acidic residues, Asp or Glu, minus the number of basic residues, Lys, Arg or His) other than 0 to 5 were defined as outside the pH range (3.7–5) of the HiRIEF strips used in this study and were discarded. Redundant sequences were collapsed into unique peptides, and their pI values predicted by the PredpI algorithm (mean absolute error = 0.026 pI units, developed in collaboration with GE Healthcare Bio-Sciences). On the basis of the predicted pI, the peptides were sorted into pI-restricted databases matching the experimental HiRIEF fractions. The pI interval of each pI-restricted database was centered at the middle pI value of the matching HiRIEF fraction, with margins based on the experimental fraction pI width and further incremented by a prediction error margin (±0.06), which was based on the prediction accuracy evaluation shown in **Supplementary Figure 9**. In the 360 databases generated (FASTA format), genomic position coordinates are coupled to each peptide sequence. The fragmentation spectra produced from each pI fraction were searched against the matching pI-restricted database using Crux (v1.37)[25] and Percolator (v2.01)[26]. A precursor mass tolerance of 10 p.p.m. and product mass tolerances of 0.02 Da for HCD-FTMS and 0.36 Da for CID-ITMS were used. Carbamidomethylation on cysteine and iTRAQ-8plex on lysine and on N termini were used as fixed modifications; and oxidation of methionine was used as variable modification.

For estimation of identification error rates, a separate decoy search was made using a pseudo-reversed version of the corresponding pI-restricted (target) database. In this decoy database, the order of all but the C-terminal residue was reversed for each peptide. All peptide-spectrum matches (PSMs) from the 72 pI fractions of each plate were pooled. From this point, two different

doi:10.1038/nmeth.2732

approaches were used as shown in **Supplementary Figure 11**. The 'global' TDA was performed using 1% FDR; and after processing steps, all single peptides (i.e., no other peptides within 10 kb) were discarded in analogy to the strategy used in a previous publication[5]. In the 'novel-only' TDA, PSMs (9–30 amino acids long) from orthogonal plates (i.e., plates corresponding to nonoverlapping pI ranges) were further pooled into groups of PSMs. These groups were considered as separate experiments, and within each group, redundant PSMs were collapsed into unique peptide identifications by exclusion of all but the highest-scoring peptide. Subsequently, all peptides already present in Ensembl (release 63, http://www.ensembl.org/), Swiss-Prot (release 2012-01, http://www.uniprot.org/) or RefSeq (release 51, http://www.ncbi.nlm.nih.gov/refseq/) were discarded. The same steps were performed on decoy PSMs, including the removal of peptides with pseudo-reversed counterparts in Ensembl, Swiss-Prot and RefSeq. Therefore, the subsequent novel-only TDA was performed on unique, previously unknown peptides. This novel-only TDA was performed using both 1% and 5% FDR at peptide level during optimization of the method, resulting in 156 and 293 novel peptides, respectively. We noticed that many of the peptides significant at 5% FDR, but not at 1% FDR, had strong support from orthogonal methods (RNA-seq, gene prediction algorithms, evolutionary conservation and additional peptides in the same locus). For instance, six of the independently validated *MYH16* peptides would have been lost using 1% FDR. Based on this, we decided to set the 5% FDR cutoff for the novel-only TDA. See also **Figure 1c** and **Supplementary Figure 11**. The resulting novel peptides were further submitted through an additional set of requirements (PSM score (Xcorr) > 2.0, and mapping to single genomic locus). Peptides that could have originated from known proteins via unusual phenomena, such as cleavage before proline residues, artifact deamidations, and N-terminal residue shuffles (wherein the first two residues swap position, a peptide-spectrum matching issue caused by the absence of the b_1 ions and the corresponding y ions in the MS2 spectrum), were discarded. The novel genomic loci mapped from these novel peptides were further populated by addition of neighboring peptides (<10 kb apart) from the global target-decoy analysis peptide list. All computations were carried out on the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) resource, provided by the Swedish National Institute for Computing (SNIC).

Novel protein-coding loci. For evaluation of the novel protein-coding loci, several different types of data were used (**Supplementary Tables 2** and **3**). Transcriptional evidence was either acquired by us (human, A431 RNA-seq data) or obtained from the public domain (mouse, C2C12 cell line, retrieved from the public domain, ENCODE/Caltech[27]). Coverage of the novel protein-coding loci by gene prediction was evaluated using three different algorithms (GenScan, N-Scan or AceView), and mammalian conservation was evaluated using the scoring algorithms (PhastCons and PhyloP), all retrieved via the UCSC Genome Browser[28] (http://genome.ucsc.edu/). Translation initiation codon type (AUG or near cognate/non-AUG) was assessed using the UCSC genome browser. To investigate whether the novel peptides could be found in cell lines other than A431, we searched publicly available MS raw data from 11 human cell lines[29] against the Ensembl 63 human protein reference database supplemented with the here-discovered 225 peptides. In addition, we compared our findings to the results of a gene prediction–dependent human proteogenomics study[11] and to suggested novel protein-coding regions from a recent study on mammalian conservation[20]. To check whether putative SNPs (single-nucleotide polymorphisms) could explain the novel peptides discovered in the human A431 cell line, we carried out analysis of the A431 mutation map[30] on the National Supercomputer Centre in Linköping (NSC). No false positives of this type were found.

The novel peptides were clustered into novel protein-coding loci according to their genomic location relative to the Ensembl gene annotation. If the locus connected to a known protein-coding gene, the locus was classified as a refined model. If the locus connected to a gene annotated as non–protein coding, the locus was classified as a pseudogene/long noncoding RNA. If the locus did not connect to any gene annotations, it was classified as intronic/intergenic.

Validation of *MYH16* novel peptides via synthetic peptides. Synthetic versions of the 20 *MYH16* peptides identified in the A431 human cell line were purchased from JPT Peptide Technologies. A mixture of the 20 peptide species was labeled by iTRAQ 8plex reagent 121, cleaned by SCX-SPE (as above), dried in a SpeedVac, dissolved in LC solvent A (final solution containing 100 ng/µl of each peptide), and analyzed by LC-MS using the same settings as described above. Annotated spectra of synthetic peptides were obtained by searching the MS raw file against a database containing only the 20 *MYH16* peptides using Sequest under PD 1.3. To obtain annotations of endogenous peptides, we did similar searches using the A431 MS raw data. The annotated MS2 spectra of synthetic peptides were then aligned to their endogenous counterparts in **Supplementary Figure 15**.

RNA sequencing and mapping. Total RNA was prepared from A431 cells (aliquots from the samples run on mass spectrometry) in biological triplicates using RNeasy Mini kit (Qiagen). Total RNA concentration was measured using a Qubit fluorometer (Invitrogen), and RNA quality was assessed using the Experion automated electrophoresis system (Bio-Rad); all samples showed high quality (RQI (RNA quality indicator) value > 9). Sequencing was performed in technical duplicates on HiSeq2000 (Illumina) using a slightly modified version of the standard Illumina RNA-seq protocol[31] with a read length of 2 × 101 bases, generating a total of 372 million paired-end reads. Mapping of the raw reads was performed using TopHat (v.1.0.14). For identification of expressed genes, estimation of FDR was done as in Ramsköld *et al.*[32]. Quantification for all human genes was obtained by calculating FPKM (fragments per kilobase of exon model per million mapped reads) values using Cufflinks (v.1.1.0), with Ensembl (release 63) as transcript reference annotation. Raw read counts were calculated (using the same transcript annotation) with HTSeq (http://www-huber.embl.de/users/anders/HTSeq/) v.0.5.1. Because pseudogene sequences often have a high number of paralogs, mapping of the RNA-seq reads to pseudogenes was done using BLAT instead of TopHat.

Data availability. MS raw data files as well as processed data files are publicly available through the repository ProteomeXchange (data set ID PXD000065). Annotations (from Proteome Discoverer)

of MS2 spectra pertaining to the novel peptides are supplied as **Supplementary Data 1**. All experimental data (both peptide and RNA-seq) pertaining to the novel gene models can be visualized within their genomic context using the UCSC genome browser (http://genome.ucsc.edu/) by loading the files and links from **Supplementary Data 2** as custom tracks. Software containing the PredpI algorithm is supplied as **Supplementary Software**.

21. Wiśniewski, J.R., Zougman, A. & Mann, M. *J. Proteome Res.* **8**, 5674–5678 (2009).
22. Reiter, L. *et al. Mol. Cell. Proteomics* **8**, 2405–2417 (2009).
23. Bjellqvist, B. *et al. Electrophoresis* **14**, 1023–1031 (1993).
24. Cargile, B.J., Sevinsky, J.R., Essader, A.S., Eu, J.P. & Stephenson, J.L. *Electrophoresis* **29**, 2768–2778 (2008).
25. Park, C.Y., Klammer, A.A., Käll, L., MacCoss, M.J. & Noble, W.S. *J. Proteome Res.* **7**, 3022–3027 (2008).
26. Käll, L., Canterbury, J.D., Weston, J., Noble, W.S. & MacCoss, M.J. *Nat. Methods* **4**, 923–925 (2007).
27. Stamatoyannopoulos, J.A. *et al. Genome Biol.* **13**, 418 (2012).
28. Kent, W.J. *et al. Genome Res.* **12**, 996–1006 (2002).
29. Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. *Mol. Cell. Proteomics* **11**, M111.014050 (2012).
30. Akan, P. *et al. Genome Med.* **4**, 86 (2012).
31. Stranneheim, H., Werne, B., Sherwood, E. & Lundeberg, J. *PLoS ONE* **6**, e21910 (2011).
32. Ramsköld, D., Wang, E.T., Burge, C.B. & Sandberg, R. *PLoS Comput. Biol.* **5**, e1000598 (2009).

doi:10.1038/nmeth.2732