Extensive Analysis of the Cytoplasmic Proteome of Human Erythrocytes Using the Peptide Ligand Library Technology and Advanced Mass Spectrometry*

Florence Roux-Dalvai‡§, Anne Gonzalez de Peredo‡§, Carolina Simó¶, Luc Guerrier¶, David Bouyssie†, Alberto Zanella**, Attilio Citterio¶, Odile Burlet-Schiltz‡, Egisto Boschetti¶, Pier Giorgio Righetti¶‖‡‡, and Bernard Monsarrat‡§§

The erythrocyte cytoplasmic proteome is composed of 98% hemoglobin; the remaining 2% is largely unexplored. Here we used a combinatorial library of hexapeptide ligands as a capturing agent to lower the signal of hemoglobin and amplify the signal of low to very low abundance proteins in the cytoplasm of human red blood cells (RBCs). Two types of hexapeptide library beads have been adopted: amino-terminal hexapeptide beads and beads in which the peptides have been further derivatized by carboxylation. The amplification of the signal of low abundance and suppression of the signal of high abundance species were fully demonstrated by two-dimensional gel maps and nano-LC-MSMS analysis. The effect of this new methodology on quantitative information also was explored. Moreover, using this approach on an LTQ-Orbitrap mass spectrometer, we could identify with high confidence as many as 1578 proteins in the cytoplasmic fraction of a highly purified preparation of RBCs, allowing a deep exploration of the classical RBC pathways as well as the identification of unexpected minor proteins. In addition, we were able to detect the presence of eight different hemoglobin chains including embryonic and newly discovered globin chains. Thus, this extensive study provides a huge data set of proteins that are present in the RBC cytoplasm that may help to better understand the biology of this simplified cell and may open the way to further studies on blood pathologies using targeted approaches. Molecular & Cellular Proteomics 7: 2254–2269, 2008.

1 The abbreviations used are: RBC, red blood cell; WBC, white blood cell; ADH, alcohol dehydrogenase; TUC, 2 m thiourea, 7 M urea, 2% CHAPS; UCA, 9 M urea, citric acid to pH 3.3; HOS, hydro-organic solution; FDR, false discovery rate; XIC, extracted ion chromatogram; Hb, hemoglobin; PK, pyruvate kinase; 2D, two-dimensional; 1D, one-dimensional; LTQ, linear trap quadrupole; IPI, International Protein Index; MFp, Mascot File Parsing and Quantification; iCIS, Interactive Chemical Information System; GO, Gene Ontology; E1, ubiquitin-activating enzyme; E2, ubiquitin carrier protein; E3, ubiquitin-protein isopeptide ligase.
glucose-6-phosphate dehydrogenase deficiency followed by (with decreasing frequency) pyruvate kinase (PK), glucose-phosphate isomerase, pyrimidine 5'-nucleotidase, triose-phosphate isomerase, and phosphofructokinase deficiencies. Yet it seems unrealistic to think that the full cytoplasmic proteome of RBCs would consist of just a handful of enzymes as known up to present times. In fact, in recent reports, 91 gene products were described by Kakhniashvili et al. (7), and an additional 252 species were identified by Pasini et al. (3) in the 2% non-hemoglobin proteome. Although these numbers are outstanding, they still fall short of expectations considering that the cytoplasmic proteomic asset would be inherited by the erythroblasts and that it is believed that living human cells should have a genetic asset of >10,000 unique gene products.

The exploration of the cytoplasmic proteome of the RBC represents a challenging analytical task as all the components present in the sample are largely masked by one heavily abundant protein, namely hemoglobin. Direct analysis of a total hemolsate by 2D gel leads to a major, smearing spot of hemoglobin with a very small number of other detectable spots (8). Thus, pretreatment of the sample is mandatory and is often performed by fractionation techniques. In different studies, the RBC cytosol was fractionated respectively by size exclusion chromatography (7), cation exchange chromatography (8), in-solution isoelectrofocusing (9), or 1D SDS-PAGE (3). Although they allow identifying a number of proteins, these generic approaches suffer from several drawbacks including limited loading capacity and inappropriate detection of the proteins co-fractionated with hemoglobin. These studies were based on standard proteomics tools, although in the study by Pasini et al. (3), sophisticated MS instrumentation was adopted that alone substantially increased the depth of proteome exploration.

We have recently described a novel approach for capturing the "hidden proteome," i.e. those rare and very rare proteins that constitute the vast majority in any cell or tissue lysate and in biological fluids (10, 11). It is based on a combinatorial library of hexameric peptide ligands bound to porous beads. Each bead contains billions of copies of a unique hexapeptide ligand distributed throughout its porous structure, and each bead potentially has a ligand different from that of every other bead. With a population of millions of individual peptide ligands obtained by combinatorial chemistry, any protein present in the starting material could theoretically interact with one or a few particular beads. Once the most abundant protein species have saturated their binding sites, the remaining molecules are washed away in the flow-through while minor protein species get progressively enriched on their corresponding beads. Thus, instead of simplifying the complex mixture into fractions or partitioning away the most abundant proteins, this approach captures most of the species present in solution up to the saturation of the solid phase ligand library and greatly reduces the dynamic concentration range of protein present in the sample. This ligand library has been efficiently used for capturing and revealing a very large population of previously undetected proteins from urine (12), serum (13), or platelets (14) as well as for "amplifying" trace impurities present in recombinant DNA products (15, 16). General reviews of this technology have also been recently published (17, 18). Here we apply this technology to the mining of the minority RBC cytoplasmic proteome and show that treatment of a hemolysate efficiently reduced the concentration of hemoglobin in the sample and unmasked numerous previously undetected proteins. We report novel methodological aspects that offer a unique insight on the performance and behavior of two combinatorial peptide libraries composed either by an amino-terminal collection of hexapeptides or by carboxylated hexapeptides. The RBC lysate was treated with both libraries, and captured proteins were analyzed on an LTQ-Orbitrap mass spectrometer. The combination of the peptide affinity-based methodology with high resolution, fast sequencing mass spectrometry allowed detection with high confidence of more than 1500 different protein species.

**EXPERIMENTAL PROCEDURES**

**Materials—** The solid phase combinatorial peptide Library-1 (ProteoMiner™), its carboxylated version (Library-2), and materials for electrophoresis such as gel plates and reagents were from Bio-Rad. N-Ethylmalamide, urea, thiourea, CHAPS, isopropanol, acetonitrile, trifluoroacetic acid, and sodium dodecyl sulfate were all from Sigma-Aldrich. Complete protease inhibitor mixture tablets and yeast alcohol dehydrogenase (ADH) were from Roche Diagnostics. Sequencing grade trypsin was from Promega (Madison, WI). All other chemical were also from Aldrich and were of analytical grade.

**Collection and Lysis of Red Blood Cells—** 80 ml of blood were collected from a healthy consenting donor by venipuncture. The blood was collected in Vacutainer tubes containing lithium heparin. The hemoglobin level from the donor was 14.2 g/dl. The blood sample was centrifuged at 1000 g at +4°C for 10 min to eliminate plasma.

Erythrocytes were washed four times with PBS + PMSF (154 mM NaCl, 10 mM phosphate buffer, pH 7.4, containing 0.1 mM PMSF). At each step supernatant and buffy coat were removed to eliminate white blood cells and the top RBC layers. Erythrocytes were filtered using an α-methylcellulose column (prepared in 10-ml syringes containing 3 ml of α-methylcellulose and 3 ml of blood). After that, an additional three washing steps were performed. At each step, along with supernatant, the upper RBC layer was removed. Filtered red cells appeared to be deprived of leukocytes. The final sample was evaluated for leukocyte (WBC) and platelet contamination by both electronic counter (LH 750 Beckman Coulter) and microscopic examination and found to be free of each of these cell types. According to the protocol used, the residual WBC content was <1 cell/1 million RBCs. The lysis of RBCs was operated by hypotonic shock. The filtered red cells were diluted to a 1:3 ratio with lysis buffer (5 mM phosphate buffer, pH 7.4, containing 1 mM EDTA and 0.5 mM PMSF) and left in ice for 30 min. At the end of the lysis step, a mixture of five protease inhibitors (Complete protease inhibitor mixture tablets, Roche Diagnostics) was added to the hemolsate. After centrifugation at 36,000 × g for 30 min at +4°C, the clear supernatant was collected. The final hemoglobin content was 5.59 g in 95 ml of hemolsate for a total protein content of 5.73 g. Hemoglobin represented 97.47% of the total protein content.
**Analysis of Red Blood Cell Lysate Fractions by SDS-PAGE**—10 μl of each sample were mixed with 10 μl of Laemmli buffer (4% SDS, 20% glycerol, 10% 2-mercaptoethanol, 0.004% bromphenol blue, 0.125 M Tris HCl, pH approximately 6.8) from Bio-Rad. The mixture was heated in boiling water for 5 min and immediately loaded in the gel. The SDS-PAGE gel was composed of a stacking gel (125 mM Tris-HCl, pH 6.8, 0.1% SDS) from Bio-Rad, pH 8.8, 0.1% SDS buffer). The cathodic and anodic compartments were filled with Tris-glycine buffer, pH 8.3, containing 0.1% SDS. The electrophoretic run was performed by setting a voltage of 100 V until the dye front reached the bottom of the gel. Staining and destaining were performed with Colloidial Coomassie Blue and 7% acetic acid in water, respectively. The SDS-PAGE gels were scanned with a Versa-Doc image system (Bio-Rad). The desired volume of each non-treated sample was diluted in Laemmli buffer and boiled for 5 min before being separated by 12% acrylamide SDS-PAGE. Proteins were visualized by Coomassie Blue staining. Each lane was cut into 20 or 22 homogenous slices that were washed with PBS until reaching the UV base line of the effluent of the second column. After the wash each individual column was subjected to three distinct elutions using, respectively, a TUC solution (2 M thiourea, 7 M urea, 2% CHAPS), a UCA solution (9 M urea, citric acid up to pH 3.3), and a hydro-organic solution (HOS) composed of 6% (v/v) acetonitrile, 12% (v/v) isopropanol, 10% (v/v) ammonia at 20%, and 72% (v/v) water. The six eluates were immediately neutralized, submitted to protein content analysis by the Bradford-Lowry standard spectrophotometric method, and then desalted by dialysis at 4 °C against a 10 mM ammonium carbonate solution (cutoff of dialysis membrane was 1000 Da) followed by lyophilization.

**Analysis of Red Blood Cell Lysate Fractions by SELDI-TOF MS**—Protein fractions at appropriate concentration, i.e. 0.02 μg/μl, were deposited upon ProteinChip® array surfaces using a Bioprocessor device. Two types of arrays were selected: CM10 (weak cation exchanger) and IMAC 30 (immobilized metal ion affinity capture) loaded with copper ions. Each array contained eight distinct spots over which the adsorption of protein could be performed. After applying the samples (the starting material and the three first eluates from each combinatorial peptide ligand library), the chip surfaces were washed to remove non-associated protein, then dried, and prepared for the analysis after application of 1 μl of energy-adsorbing matrix solution composed of a saturated solution of sinapinic acid in 50% acetonitrile and 0.5% trifluoroacetic acid. All arrays were then analyzed with a PCC 4000 ProteinChip MS reader. The instrument was used in a positive ion mode with an ion acceleration potential of 20 kV and a detector gain voltage of 2 kV. The mass range investigated was from 3 to 20 kDa. The laser intensity was set between 200 and 250 units according to the sample tested. The instrument was mass-calibrated with a kit of a standard mass mixture, “All-in-1 protein standard.”

**2D PAGE Analysis**—The desired volume of each non-treated sample and eluates was solubilized in the “2D sample buffer” (7 M urea, 2 M thiourea, 3% CHAPS, 40 mM Tris, 5 mM tributyolphosphate, 10 mM acrylamide) to a final concentration of 2 mg/ml protein, and the alkylation reaction was allowed to proceed at room temperature for 30 min. To stop the alkylation reaction, 10 mM DTT (diluted from a bottle of neat DTT) was added to the solution followed by 0.5% Ammonium (diluted directly from the stock, 40% Ammonium solution) and a trace amount of bromphenol blue. 18-cm-long IPG strips (Bio-Rad), pH 3–10, were rehydrated with 400 μl of protein solution for 5 h. IEF was carried out with a Protean IEF cell (Bio-Rad) in a linear voltage gradient from 100 to 2000 V for 5 h and 2000 V for 4 h followed by an exponential gradient up to 10,000 V until each strip was electrophoresed for 25 kV-h. For the second dimension, the IPG strips were equilibrated for 25 min in a solution containing 6 μl urea, 2% SDS, 20% glycerol, 375 mM Tris-HCl (pH 8.8) under gentle shaking. The IPG strips were then laid on an 8–18% acrylamide gradient SDS-PAGE gel with 0.5% agarose in the cathode buffer (192 mM glycine, 0.1% SDS, Tris to pH 8.3). The electrophoretic run was performed by setting a current of 5 mA/gel for 1 h followed by 10 mA/gel for 1 h and 15 mA/gel until the dye front reached the bottom of the gel. Gels were incubated in a fixing solution containing 40% methanol and 7% acetic acid for 1 h followed by silver staining. Destaining was performed in 7% acetic acid until the background was clear followed by a rinse in pure water. The two-dimensional electrophoresis gels were scanned with a Versa-Doc image system (Bio-Rad) by fixing the acquisition time at 10 s; the relative gel images were evaluated using PDQuest software (Bio-Rad). After filtering the gel images to remove the background, spots were automatically detected, manually edited, and then counted.

**LC-MS/MS as Analytical Method for the Identification of Proteins in Lysate Fractions**—100 μg of each elution fraction as well as 2000 μg of the initial non-treated RBC lysate were diluted in Laemmli buffer and boiled for 5 min before being separated by 12% acrylamide SDS-PAGE. Proteins were visualized by Coomassie Blue staining. Each lane was cut into 20 or 22 homogenous slices that were washed in 100 mM ammonium bicarbonate for 15 min at 37 °C followed by a second wash in 100 mM ammonium bicarbonate, acetonitrile (1:1) for 15 min at 37 °C. Reduction and alkylation of cysteines were performed by mixing the gel pieces in 10 mM DTT for 35 min at 56 °C followed by 55 mM iodoacetamide for 30 min at room temperature in the dark. An additional cycle of washes in ammonium bicarbonate and ammonium bicarbonate/acetonitrile was then performed. Proteins were digested by incubating each gel slice with 0.6 μg of modified sequencing grade trypsin (Promega) in 50 mM ammonium bicarbonate overnight at 37 °C. The resulting peptides were extracted from the gel by three steps: a first incubation in 50 mM ammonium bicarbonate for 15 min at 37 °C and two incubations in 10% formic acid, acetonitrile (1:1) for 15 min at 37 °C. The three collected extractions were pooled with the initial digestion supernatant, dried in a SpeedVac, and resuspended with 14 μl of 5% acetonitrile, 0.05% trifluoroacetic acid.

The peptide mixtures were analyzed by nano-LC-MS/MS using an Ultimate3000 system (Dionex, Amsterdam, The Netherlands) coupled to an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). 5 μl of each sample (for the slices at hemoglobin size in the non-treated sample, the sample was 2× diluted) were loaded on a C18 precolumn (300-μm inner diameter × 5 mm, Dionex) at 20 μl/min in 5% acetonitrile, 0.05% trifluoroacetic acid. After 5-min desalting, the precolumn was switched on line with the analytical column (75-μm inner diameter × 15 cm, PepMap C18, Dionex) equilibrated in 95% solvent A (5% acetonitrile, 0.2% formic acid) and 5% solvent B (80% acetonitrile, 0.2% formic acid). Peptides were eluted using a 5–50% gradient of solvent B during 80 min at 300 nL/min flow rate. The LTQ-Orbitrap was operated in data-dependent acquisition mode with the Xcalibur software. Survey scan mass spectra were acquired in the Orbitrap in the 300–2000 m/z range with the resolution set to a value of 60,000. The five most intense ions per survey scan were selected for CID fragmentation, and the resulting fragments were analyzed in the linear trap (LTQ). Dynamic exclusion was used within 60 s to prevent repetitive selection of the same peptide.

**Database Search and Data Analysis**—The Mascot Daemon software (version 2.2.0, Matrix Science, London, UK) was used to perform database searches in batch mode with all the raw files acquired on each gel lane. To automatically extract peak lists from Xcalibur raw files, the ExtractMSN macro provided with Xcalibur (version 2.0 SR2,
Thermo Fisher Scientific was used through the Mascot Daemon interface. The following parameters were set for creation of the peak lists: parent ions in the mass range 400–4500, no grouping of MS/MS scans, and threshold at 1000. A peak list was created for each fraction analyzed (i.e., each gel slice), and individual Mascot searches were performed for each fraction. Data were searched against all entries in the Internation Protein Index (IPi) human v3.34 protein database (67,764 sequences). Carbamidomethylation of cysteines was set as a fixed modification and oxidation of methionines and protein amino-terminal acetylation were set as variable modifications for all Mascot searches. Specificity of trypsin digestion was set for cleavage after Lys or Arg except before Pro, and one missed trypsin cleavage site was allowed. The mass tolerances in MS and MS/MS were set to 5 ppm and 0.6 Da, respectively, and the instrument setting was specified as “ESI-Trap.” Mascot results were parsed with the in-house developed software Mascot File Parsing and Quantification (MFPaQ) version 4.0 (19), and protein hits were automatically validated if they satisfied one of the following criteria: identification with at least one top ranking peptide with a Mascot score of more than 41 (p value < 0.001) or at least two top ranking peptides each with a Mascot score of more than 24 (p value < 0.05). When several proteins matched exactly the same set of peptides, only one member of the protein group was reported in the final protein lists in supplemental data 1 and 2 for more clarity (the one returned by Mascot in the protein summary list). However, detailed protein groups are shown in supplemental data 3 and can be viewed at http://proteomique.ipbs.fr/rcb through the MFPaQ interface using the Mozilla Firefox Web browser. Moreover in each Mascot result file, the MFPaQ software detected highly homologous Mascot protein hits, i.e., proteins identified with top ranking MS/MS queries also assigned to another protein hit of higher score (i.e., red, non-bold peptides). These homologous protein hits were validated and included in the final list only if they were additionally assigned a specific top ranking (red and bold) peptide of score higher than 41. Otherwise they were considered as proteins matching a subset of peptides from another hit and were eliminated from the final lists shown in supplemental data 1, 2, and 3 (however, these homologous entries can be viewed at http://proteomique.ipbs.fr/rcb through the MFPaQ interface where they are indicated in italic). From all the validated result files corresponding to the fractions of a 1D gel lane, MFPaQ was used to generate a unique, non-redundant list of proteins found in different gel slices by creating clusters of protein groups (composed of all the protein sequences matching the same set of peptides) if they have at least one common member. Global protein lists corresponding to the merging of several protein lists (for example, the three different eluates from one peptide library column corresponding to three different 1D gel lanes) were generated in the same way to eliminate redundancy. Protein list comparisons were also based on the comparison of protein groups (hits matching the same set of peptides) from different lists, and the software assigned these protein groups as “shared” or “specific” depending on whether or not they have common members. The contaminant proteins, i.e., desmplakin, desmoglein-1, desmocollin-1, and all keratins, were manually removed from the protein lists. To evaluate the false positive rate in these large scale experiments, all the initial database searches were performed using the “decoy” option of Mascot, i.e., the data were searched against a combined database containing the real specified protein sequences (target database, IPi human) and the corresponding reversed protein sequences (decoy database). MFPaQ used the same criteria to validate decoy and target hits, computed the false discovery rate (FDR = number of validated decoy hits/(number of validated target hits + number of validated decoy hits) × 100) for each gel slice analyzed, and calculated the average FDR for all slices belonging to the same gel lane (i.e., to the same sample). The FDR found for the analysis of the non-treated sample and of the six different eluates after peptide library treatment ranked between 0.33 and 1.68% with an average of 1.16%.

Reproducibility of LC-MS/MS Analysis—100 μg of UCA eluate of Library-1 were separated by 12% SDS-PAGE, and 20 gel slices were cut and treated as described above. After peptide extraction and drying, the samples were resuspended in 17 μl of 5% acetonitrile, 0.05% trifluoroacetic acid. The 20 samples were successively analyzed on the nano-LC-MS/MS system, and this analysis was made in triplicate. Mascot was used for database searches, and MFPaQ was used for protein validation, generation of protein lists, and comparison of these lists as described above.

Nano-LC/MSMS Quantitative Experiment after Peptide Library Treatment—100, 300, or 1000 pmol of yeast alcohol dehydrogenase were added to 680 mg of red blood cell lysate prior to treatment with a 0.5-mL library-1 column. The column was then washed as described above and eluted with UCA buffer. Three replicate treatments were performed for each ADH amount. 125 μg of each of the nine eluates were separated by 12% SDS-PAGE, and a slice was cut in each lane at the ADH molecular mass. The gel slices were treated as described above, and samples were resuspended in 17 μl of 5% acetonitrile, 0.05% trifluoroacetic acid and analyzed three times by nano-LC-MS/MS with the same method described above. Xcalibur software was used to perform the ion current chromatogram extraction of ADH peptides identified by MS/MS with a mass tolerance of 5 ppm and to perform the area integration of the corresponding peaks using the ICIS algorithm. To estimate the abundance of the protein in the different samples, the extracted ion chromatograms (XIC) was integrated for eight parent ions assigned to ADH, summed area values were calculated from these different ion peaks, and average values were then calculated from the three different MS injections for the same sample.

RESULTS

Treatment of Red Blood Cell Proteins and Preliminary Analysis—A total amount of 5730 mg of RBC cytosolic proteins was loaded on a sequence of two columns composed of the two different ligand libraries, and the captured proteins were collected by eluting each column sequentially with TUC, UCA, and HOS buffers. From the six resulting eluates, 8.14 mg of protein were collected in total, representing about 0.15% of the initial input. For the first library column, TUC, UCA, and HOS eluates contained, respectively, 1.52, 5.51, and 0.19 mg of protein, and for the second library column, they contained 0.13, 0.77, and 0.2 mg of protein. SDS-PAGE analysis of all fractions under reducing conditions revealed different patterns of proteins; as shown in Fig. 1A, hemoglobin, the most abundant protein (lanes “Non-treated” and flow-through “FT”), was considerably diminished in the eluates, whereas many other proteins appeared, covering a large mass interval from 10 up to 250 kDa. The richest fraction was the UCA eluate from Library-1. Fractions containing the least amount of proteins were hydro-organic eluates from both Library-1 and Library-2. In Fig. 1B an analysis performed using SELDI mass spectrometry shows the distribution pattern for polypeptides of relatively small masses and confirms a large number of complementary properties from Library-1 and Library-2 as well as from the various eluates. Spectra of UCA elutions of both libraries show the presence of a large number of species that are absent from the non-treated sample.
Two-dimensional maps were also run for all six eluates, revealing in all cases several hundred protein spots as compared with a handful from the control RBC lysate. Instead of presenting all maps of the individual fractions, Fig. 2 shows the first two eluates from each column (TUC and UCA eluates from Library-1 and Library-2) and also the control gel compared with a protein mixture composed of all six eluates. One can appreciate at a glance the drastic change of landscape: in the control map (upper rightmost
hemoglobin from the initial crude extract, clearly visible in the lower molecular mass region as denatured \( \alpha \) and \( \beta \) chains including some oligomeric globin aggregates that are mostly dimers and trimers, constitutes 98% of the entire proteome. From the remaining 2% one can scarcely count some 80 spots. Conversely the map with the mixture of all eluates (lower rightmost panel) appears to be fully covered by spots, approximately 950 via PDQuest count, covering the entire pI 3–10 and the molecular mass interval from 10 to 250 kDa, and this notwithstanding the fact that in the latter half as much protein (640 \( \mu \)g) had been loaded as compared with the control (1300 \( \mu \)g). The comparison between TUC eluates shows a relatively limited number of differences: although the eluate from Library-1 appears more populated with proteins, most of them are located within similar areas. Conversely the comparison of UCA eluates clearly shows two types of areas covered by protein spots: dominantly acidic with medium-high masses in the eluate from Library-1 and mostly alkaline with medium-low masses in the eluate from Library-2.

**Protein Identification by Nano-LC-MS/MS**—To go further in the characterization of the cytoplasmic RBC proteome, the proteins from the two library columns were analyzed by nano-LC-MS/MS after fractionation by 1D SDS-PAGE. In principle, the six eluates could have been combined and analyzed in a single experiment. However, we performed a separate analysis for each of the eluates to (i) maximize the number of proteins identified in that large scale study by giving the mass spectrometer more time to acquire more MS/MS spectra and (ii) get a detailed composition of the individual eluates and provide a better understanding of the behavior of the two peptide library columns. Moreover to evaluate whether the peptide library treatment provided a real advantage in terms of number of proteins identified by nano-LC-MS/MS, we also performed the analysis on the crude RBC lysate before any treatment. For the six eluates from the two libraries, 100 \( \mu \)g of protein were loaded on the SDS-PAGE gel. On the other hand, for the crude RBC lysate, we massively overloaded the SDS-PAGE gel with up to 2 mg of starting material to have better chances to detect proteins other than hemoglobin. Although we obtained a huge band of hemoglobin on the 1D gel and a quite distorted migration lane, we could also enhance the staining of less abundant protein species migrating in the upper region of the gel and maximize the number of proteins identified in that region. Thus, by doing that, we could really compare the results obtained using the peptide libraries method with those of a direct, straightforward fractionation approach at the limit of the loading capacity. In all cases, after the electrophoretic run, each track was cut into 20 slices along the migration path, and the protein content was digested by trypsin and subjected to MS/MS analysis on the LTQ-Orbitrap mass spectrometer. Fig. 3 gives a summary of the identifications for the different analyses. The non-treated RBC lysate revealed 535 gene products (for comparison, the same analysis performed with 100 \( \mu \)g of non-treated RBC lysate instead of 2 mg yielded only 331 identifications; data not shown). On the other hand, analysis of 100 \( \mu \)g of the six eluates from the library beads led to a global number of 1524 proteins identified (534 exclusive to Library-1, 254 exclusive to Library-2, and 736 in common) (Fig. 3B). There was a large overlap with the set of proteins identified in the non-treated sample as 470 of the 535 initially identified proteins were found again in the eluates. On the other hand, as many as
1052 proteins were identified only in the eluates (Fig. 3A). A detailed concatenated list corresponding to all the proteins identified in that study (non-treated RBC cytoplasm plus all six eluates), i.e. 1578 non-redundant proteins, is given in supplemental data 1. The number of proteins identified in each eluate (Fig. 3, C for Library-1 and D for Library-2) varies between the two libraries. However, it can be seen that for almost all the eluates analyzed, more proteins could be identified from 100 μg of material than the 535 species identified from 2 mg of crude lysate. The richest eluate was the UCA fraction from Library-1, which alone allowed detection of more than a thousand proteins. A more in-depth analysis of the effectiveness of eluting agents reveals that for the first library the UCA elution was the most efficient, whereas for the second library, more species were detected in the elution performed with normal TUC buffer. For each library, the last elution with HOS buffer contributed to detect an additional 128 and 73 species, respectively, representing about 11 and 8% of the total. Finally the comparison between the two libraries allows assessment of the significant contribution of Library-2 compared with Library-1 (regular ProteoMiner beads). A substantial overlap was found between the species captured by the two types of columns (736 proteins common to the two libraries). However, the use of the second peptide library column connected on line with the first peptide library column still allowed the addition of 254 novel proteins to the final count (Fig. 3B). Thus, the carboxylated column adds about 16% more species to the global protein discovery.

It must be noted that this number, as well as all the other overlaps shown in Fig. 3, results from automatic protein list comparison. It is well known, however, that this kind of comparison can be unreliable in cases of large undersampling of the mass spectrometer, i.e. when the number of ions detected in MS mode is too high for the duty cycle of the machine so that not all of them can be picked up for MS/MS. In such cases, the reproducibility of the protein lists obtained is quite poor. It has been reported that for complex mixtures twice injecting the same sample may lead to identification of as much as 34% new proteins in the second run compared with the first run as not exactly the same ions are selected for MS/MS (20). To assess the reproducibility of our analysis and check whether protein list comparison was relevant, we analyzed three times the UCA fraction from Library-1 by the same 1D gel LC-MS/MS method as described above. In doing that, we could identify, respectively, 1070, 1053, and 1031 proteins and found that the second analysis brought only 4.9% new proteins compared with the first analysis, whereas the third analysis just contributed to add 3.1% new species to the total. This shows that the reproducibility of the protein lists we obtained is quite good, around 95%, even for the most complex of the eluates that we analyzed. Thus, the overlaps calculated in Fig. 3 should reflect a real situation although with about a 5% error margin.

**Effect of the Combinatorial Library Treatment on Nano-LC-MS/MS Analytical Coverage**—The numbers given above show that treatment of the sample with the two peptide libraries greatly enhanced the number of proteins identified by nano-LC-MS/MS; this must result from a better MS/MS sampling of the ions detected. To illustrate that effect, we compared the MS/MS queries assigned to identified proteins in the sample before treatment versus the UCA eluate from Library-1. A total number of 429 proteins were identified in common between these two samples and are represented in the histogram in Fig. 4 from high to low Mascot score, reflecting protein abundance (decreasing abundance, from left to right). In the upper part of the graph, the number of MS/MS spectra performed for each protein from the non-treated RBC lysate sample is plotted, whereas in the lower part, the corresponding number of MS/MS spectra is plotted for the same proteins identified in the UCA eluate. For the non-treated sample, there is an exponential decay of the number of MS/MS spectra when going toward the region of low abundance proteins, reflecting the large dynamic range of the sample. Conversely for the UCA eluate, although some proteins in the high abundance region still show high values of MS/MS, such values remain at a much higher level even in the low abundance region. This illustrates the effect of peptide ligand library treatment on the concentration of proteins in the cell lysate; this can be further appreciated by looking at the sequence coverage obtained for two particular proteins, one in the high abundance region and one in the low abundance region (Fig. 4, A and B). These panels show tables summarizing the number of peptide sequences and number of MS/MS queries that were assigned to the protein. For a very abundant protein like peroxiredoxin 2 (A), 25 unique peptides were identified by Mascot, giving a sequence coverage of 89% in the case of non-treated sample. However, to achieve this result as many as 1018 MS/MS spectra were acquired during the analysis of the gel slice where this protein was most abundant. Clearly very abundant ions from this protein were eluted continuously during the nano-LC run, probably suppressing the signal of lower abundance species, and were repeatedly sequenced by the mass spectrometer. Conversely in the eluate from the first library column, Mascot identified 14 peptides for this protein (73% sequence coverage), but “only” 71 MS/MS spectra were acquired during the run. Thus, by avoiding acquiring too much redundant information on that protein, the mass spectrometer could spend more time on less abundant species. This effect can be seen on a minor protein like pyrimidine 5’-nucleotidase, an important enzyme involved in the catabolism of nucleotides, whose deficiency in RBCs is often associated with hereditary non-spherocytic hemolytic anemia. Although this protein was barely identified in the starting material with three MS/MS spectra and three peptides, it was very well detected in the eluate from the column with 48 MS/MS spectra and 24 peptides. Thus, as can be seen on the global histogram, less time was spent on abundant species, and more time was
spend on minor species. Moreover, the treatment enabled identification of 625 additional proteins that were not detected at all in the starting material either because the signal of the corresponding ions was suppressed or because the mass spectrometer missed them for MS/MS sequencing.

**Effect of the RBC Lysate Treatment on Quantitative Information**—As shown above, the peptide library treatment greatly affects the number of MS/MS per protein, which can be used as a semiquantitative indicator of the abundances of proteins in a complex mixture (21). Indeed after the treatment of the sample, Fig. 4 shows that the concentration of a protein can be either increased or decreased and that both effects can have variable amplitude depending on the protein considered. This is probably because the degree of affinity for a particular bead is intrinsically protein-dependent. The next question then is whether this method can still retain relative quantitative information and could be used for example in the context of differential studies on RBC samples. To check this point, we performed a spiking experiment at three different concentrations with a purified yeast protein (ADH) in the RBC lysate and checked the respective ratios found for this protein in the three samples before and after treatment with the peptide library. Either 100, 300, or 1000 pmol of ADH were added in 8 ml of RBC lysate (680 mg of protein in total), and the samples were treated with the standard library column. Each spiking experiment was performed in triplicate, and
each treatment with library beads was also performed in triplicate. The eluates from the columns were then fractionated by 1D SDS-PAGE, and for all lanes, a gel slice corresponding to the molecular mass of ADH was analyzed by nano-LC-MS/MS. Each nano-LC-MS/MS analysis was performed in triplicate as well, resulting finally in 27 different measurements. For all the spikes ADH could be detected within the massive presence of RBC proteins with a number of assigned peptides typically ranging from 3 (100-pmol spike) to 15 (1000-pmol spike). The abundance of the protein in the different samples was estimated by extracting the ion current chromatogram for eight parent ions assigned to ADH and integrating the area under each elution peak. Summed area values could be obtained from these different ion peaks, and average values were then calculated from the different MS injections for the same sample. It can be seen in Fig. 5 that the MS signal clearly goes up when the amount of protein spiked in the lysate increases as shown on an individual parent ion assigned to ADH (Fig. 5A) or on the global average value calculated from all the different peptides and MS injections (Fig. 5B). Moreover this value shows very little variation between the different replicate experiments, demonstrating the good reproducibility of the peptide library treatment. When calculating a ratio from these values between the respective samples (300 versus 100 pmol and 1000 versus 100 pmol), numbers very close to the expected values of 3 and 10 were obtained (Fig. 5C). Thus, the relative abundance ratio existing for a protein between two different untreated samples ap-
pears to be unchanged after peptide library treatment of the two samples.

Functional Annotation of the Proteins Identified in the RBC—To go further in the analysis of the list of proteins identified in the RBC cytoplasm, we used automatic annotation tools like GoMiner High-Throughput and the Ingenuity Pathway Analysis (Ingenuity Systems, Redwood City, CA) software for data interpretation.

Regarding subcellular location, a majority of proteins were categorized as cytoplasmic by the two softwares. For example, out of 535 proteins identified in the non-treated material, 348 were annotated with a GO term related to cellular localization, and Gominer classified 72% of these annotated proteins as cytoplasmic. However, a relatively high number of proteins were also classified as nuclear, i.e. 28% in the non-treated RBC lysate and up to 33% in the treated sample. It must be noted that some overlap may exist between these categories as many proteins assigned with the GO term "nuclear" can also have a possible localization in the cytoplasm. Although the RBCs are enucleated when they enter the blood circulation, the presence of proteins classified as nuclear probably reflects the fact that a certain number of proteins that normally stay in the nucleus in the erythroblasts are still present in the RBC cytosol.

GoMiner and Ingenuity were also used to perform a functional classification of the proteins identified in this study. This was done on the lists obtained either before or after treatment of the sample. The two softwares assign the proteins to different functional classes based either on the annotation of protein entries with GO terms in protein databases (GoMiner) or on the data contained in a proprietary, in-house curated knowledge database (Ingenuity). The latter tool can also assign the proteins to specific, well defined canonical pathways related to various biological processes that are contained in the knowledge database. Both programs provide an "enrichment" or "ratio" value measuring the extent to which a functional class or pathway is represented in the list of interest compared with a background reference. In the case of GoMiner analysis, we provided a reference list constituted by the entire set of protein entries from the IPI human database. Thus, the software compares the number of proteins from the list assigned in a particular category with the number of proteins in the IPI database assigned to the same category. A p value and a false discovery rate reflecting the statistical significance for enrichment of categories are also computed. In the case of Ingenuity analysis, ratios and p values are automatically computed both for functional classes and for canonical pathways. Fig. 6 shows the best represented canonical pathways in the non-treated RBC lysate, ranked by decreasing ratio values, and their corresponding ratios in the treated sample as calculated by Ingenuity. Most of them seem relevant in light of the known function of the RBC and of the known biological processes taking place in that cell: degradation of proteins (protein ubiquitination pathway), degradation of nucleic acids (purine and pyrimidine metabolism pathways), anaerobic glycolysis (pentose phosphate and
glycolysis pathways), and protection against oxidative damages because of the high content of oxygen in the cell (oxidative stress response and glutathione metabolism pathways). These major processes of the RBC give high ratio values for the related pathways in the non-treated as well as in the treated samples. However, some unexpected pathways are also highlighted by the software in the case of the treated sample, like various signaling pathways. This may be due to the fact that many minor species were identified in the treated RBC lysate. For example, several mitogen-activated protein kinases or mitogen-activated protein kinase (MAPK)-interacting proteins were found in the treated RBC lysate (eight versus three in the untreated sample), accounting for the high scoring attributed to signaling pathways by Ingenuity. Although these signaling cascades may no longer be active in mature RBCs, several molecules taking part in them are still present in the cell, and their identification after treatment of the sample with peptide libraries illustrates the in-depth characterization of the proteome allowed by the described method. Comparable results were obtained when performing functional annotation of the protein lists using the GoMiner software. The list of GO terms and their number of associated proteins along with their enrichment value, p value, and FDR computed by GoMiner for each protein list is given in supplemental data 5. Among the best enriched GO biological processes, many GO categories corroborating the major expected function of RBCs can be found either in treated or non-treated samples (e.g. "gas transport," "heme metabolic process," "glycolysis," "oxygen and reactive oxygen species metabolism," etc . . . ). However, in the case of treated RBC lysate, some unexpected GO terms appear as well ranked in the list, although they may not reflect a real process taking place in RBCs. Again this indicates that many low abundance species, probably inherited from erythroblasts, are still present in RBCs and could be identified thanks to the peptide libraries beads.

**DISCUSSION**

**Performance of the Peptide Library Treatment for the Analysis of the RBC Cytoplasmic Proteome**—In this study, we analyzed the cytoplasmic fraction obtained from highly purified RBCs and treated these soluble proteins with different combinatorial libraries of hexapeptide ligands to reduce the dynamic concentration range of the sample and obtain access to the identification of minor species. Three different eluates from each peptide library column as well as an untreated sample of cytoplasmic proteins were analyzed by 1D gel fractionation and nano-LC-MS/MS on an LTQ-Orbitrap mass spectrometer, leading to the identification of 1578 different species. This total list of cytoplasmic proteins identified in the RBC cell sap is surprisingly large given that RBCs are simplified cells without nucleus and organelles and considering previous reports in the literature. Indeed in a recent study performed with an LTQ-FT mass spectrometer, Pasini et al. (3) extensively analyzed using various extraction protocols the membrane fraction of RBCs as well as the soluble fraction in which they reported the identification of 252 proteins. When analyzing a comparable sample (i.e. untreated, cytoplasmic soluble fraction), we could identify 536 gene products, but this was obtained by applying a huge amount of sample to the SDS-PAGE gel (2 mg of total protein, grossly overloading an analytical strip). When we applied the regular amount loaded in typical studies (100 µg) we saw with the LTQ-Orbitrap essentially the same number of proteins (approximately 300) as in Pasini et al. (3) (data not shown). The proteins identified in our study were automatically validated with relatively stringent filtering rules leading to a false discovery rate of about 1% as estimated by target-decoy database searches, and redundancy was eliminated with the in-house developed software MFPaQ (19). Thus, our final validated and non-redundant protein list should reflect the real complexity of the sample analyzed. Regarding the purity of the RBC preparation, the contamination with leukocytes (WBCs) can be estimated to be less than one WBC per one million RBCs, and no contamination of the sample by serum proteins or by membranous species carried over during the lysis of RBCs could be inferred from analysis of the list of proteins identified. For example, complement protein C4, which is present in the blood stream and is known to coat the RBC membrane, is not found in our list. Thus, the proteins identified here should be genuine components of the cytoplasm from a homogeneous population of erythroid circulating cells composed in great majority of mature RBCs and of a few reticulocytes (typically 0.5–1%). In fact, the major reason that accounts for the high total number of protein species reported here, as compared with previous studies, rests most probably on the application of the hexapeptide ligand library technology, namely the ability of sharply cutting the high abundance proteins while greatly enriching the low abundance species until full saturation of the corresponding ligand in the library. This reduction of dynamic concentration range effect was clearly shown in our study by various techniques (1D and 2D gels, SELDI, and nano-LC-MS/MS), each of them demonstrating the fact that many minor species became detectable after treatment of the sample. For example, more than a thousand proteins were identified from only 100 µg of the UCA eluate from the first peptide library column to be compared with the 535 gene products identified from 2 mg of untreated sample, an amount probably representing the limit of material that can be applied for classical 1D gel fractionation.

To increase the analytical coverage of body fluids characterized by a large dynamic concentration range of proteins, many strategies, mostly based on sample prefractionation or on immunodepletion of major species, have been extensively described. Although prefractionation tools based on general physicochemical parameters of proteins (size, charge, and hydrophobicity) may be useful to segregate some abundant proteins in specific fractions, thus enhancing the detectability of minor species in other fractions, they often suffer from
several limitations such as implementation difficulties, limited loading capacity, questionable reproducibility, and limited selectivity power. This last point implies that highly abundant species often co-fractionate with many other minor proteins that will thus still be undetectable. In the case of immunosubtraction methods, which are based on specific antibodies raised against the most abundant proteins of the sample, a major reported drawback is probably the co-depletion and loss of many proteins either bound to the target protein to be subtracted or in some way recognizing the capturing antibody (22). In our experience, using the approach described here, very few proteins are lost while the relative concentration of minor species is increased. One limitation of the approach is to load enough material on the peptide library column to saturate the binding sites of abundant proteins. This may not be possible in clinical studies dealing with limited amounts of biological samples, and further improvements will have to be made to scale down the peptide library bead treatment in such cases. However, when the fluid is easily available, huge amounts of sample can be processed (5.6 g of protein from RBC lysate in the present study), further incrementing the enrichment of the final sample in low abundance proteins in a quite reproducible manner (see below).

Using this method on the RBC lysate, we could show here that a better MS/MS sampling of the ions from minor species was obtained as a result of the reduction of protein concentration dynamic range. In a recent study, de Godoy et al. (23) examined the different parameters that affect the analytic coverage of complex mixtures by nano-LC-MS/MS. They report that the major limitation is not the sensitivity of the mass spectrometer but rather a combination of effective sequencing speed (the mass spectrometer is not fast enough to sequence all the ions detected) and dynamic range (high abundance peptides prevent detection of low abundance peptides because of signal suppression). The use of sample treatment with peptide library beads clearly alleviates these instrumental limitations and allows globally more MS/MS on a higher number of species. It appears that after treatment of the sample the limitation due to sequencing speed seems no longer to be a crucial issue. Indeed reproducibility of the protein lists obtained was shown to be about 95%, a number significantly higher than typical values obtained for analysis of complex proteomes by nano-LC-MS/MS, suggesting that almost complete MS/MS sampling of the ions detected was performed and that nearly everything that was detected was probably also sequenced. It is possible, however, that some proteins of the sample were not identified either because they did not find a good ligand in the combinatorial peptide libraries and did not bind efficiently to the beads (this is probably reflected in the fact that 56 proteins identified in the non-treated sample could not be found after treatment) or because they were too dilute in the initial sample and the enrichment provided by the method was still insufficient to bring the concentration of these species into the 10^4 dynamic range typically explored by the mass spectrometer. However, the fact that very minor species, never detected before, could be identified in this study (discussed below) suggests that the list of RBC cytoplasmic proteins provided here is probably as exhaustive as possible with current technology.

Normal Peptide Library Versus COOH-modified Peptide Library—The use of two peptide libraries was justified by their complementary behavior already described in the case of platelet extracts (14) and bile fluid protein discovery (24). Notwithstanding the large protein overlap, the use of the second library always allowed detection of additional proteins that were not captured by the first set of peptides (16% more in the present case). Clearly the different behavior is attributed to the difference of the terminal group of the peptides chains. The presence of weak cationic (primary amine) or anionic (carboxylic group) character on the terminal group may induce a difference of affinity between the peptide chain and the bait protein because of the modification of the dissociation constant between the peptide and the partner proteins present in the initial crude lysate. Generally peptides with carboxylic terminal groups interact more efficiently with cationic proteins, whereas amino-terminal peptides interact mostly with anionic proteins. This behavior is illustrated by the 2D gels shown in Fig. 2, panels UCA, Library-1 and UCA, Library-2. Although Library-1 contributed to a larger number of exclusively captured species (see Fig. 3B), it did not have all necessary baits to capture all proteins from the crude extract. 254 additional proteins were found thanks to the use of Library-2. This represents about 16% of proteins that would not have been detected without the use of the second library.

Interestingly when observing the amount of proteins eluted by the sequence of desorbing agents, UCA was the most efficient desorbing solution for Library-1. A possible explanation is that when using citric acid, a trivalent compound bearing three negative charges, such a negative surface would efficiently compete for the amino binding sites of the hexapeptides on the beads, thus helping to discharge all acidic proteins by breaking the ionic bonds while the hydrogen bonds would be loosened by urea. Moreover it can be seen on the 2D gel maps that the UCA eluates from the two libraries present an overlap that is significantly lower than that for the TUC eluates. Clearly acidic proteins are preferentially eluted by UCA in the case of Library-1. The HOS eluent, a hydro-organic mixture designed for the desorption of highly hydrophobic proteins, allowed collection of only a minor amount of proteins but contributed to enlarge the list of discovered species by 201 additional gene products. This relatively modest number is not a real surprise because all proteins identified in the present study were from the soluble part of the red blood cell.

Quantitative Information—Treatment with peptide ligand libraries is intrinsically an approach that will modify the abundances of proteins in the treated versus initial sample. Highly
The Unknown Cytoplasmic Proteome of Human Erythrocytes

abundant species should quickly saturate their partner baits on the beads, and their final concentration in the treated sample will be greatly decreased. On the other hand, the enrichment factor of other species will depend on the affinity of a given protein toward its recognized ligand, and the order of abundance of diverse proteins in the final sample may be changed. It was thus important to assess whether or not the method could be used for differential proteomics studies, i.e. whether proteins differentially expressed in samples to be compared are still found with the same differential expression ratio after treatment of these samples. In other words, although the enrichment -fold during the treatment may vary from protein to protein in the same sample, it should be reproducible for a given protein present in different samples treated in parallel. To assess that, different amounts of a standard protein were spiked in samples of RBC lysate, and replicate peptide library treatments were performed on these samples. We show here that the treatment seems to be very reproducible as indicated by the XIC area measurements of the peptide ions corresponding to the spiked protein. Indeed for a given amount of this protein, the standard deviation of the values measured for different replicate treatments was lower than the standard deviation obtained for the different MS measurements themselves performed on the same sample. Moreover for different amounts of the spiked protein, the theoretical initial ratio was found to be nearly the same after treatment of the samples. Thus, as long as the protein does not saturate the beads, the relative quantitative information on this protein between different samples is well conserved during the treatment. This will need to be confirmed on a larger number of proteins, and a more precise evaluation of the number of proteins that effectively saturate the beads and for which the relative quantitative information should be lost (at least in the case of label-free approaches) will have to be performed. However, our preliminary experiment, as well as the successful identification of several specific differentially expressed species in proteomics analysis using this methodology, suggests that treatments with peptide libraries can be used in the context of differential studies, for example on RBC pathologies or for biomarker discovery in body fluids.

Biological Significance of the Proteins Identified—Many of the proteins identified in this study are well known components of the RBC cytoplasm that accomplish important functions in this cell. Among them are of course all the hemoglobin chains but also the proteins involved in essential metabolic functions of the RBC such as glycolysis or protection against oxidative damage described in more detail below. However, more unexpected classes of biological processes or cellular localization were also found when automatic annotation of the list was performed with software like Ingenuity Pathway Analysis or GoMiner. Considering the high number of proteins identified, it is likely that some of them are inherited from the erythroblast but are present in the RBC in an inactive state. Differentiation of RBCs occurs in well defined stages in the bone marrow from proerythroblasts to early erythroblasts and then late erythroblasts. In these nucleated intermediates, many pathways are activated to establish the differentiation program in response to microenvironment stimulation and particularly erythropoietin signaling. This promotes adequate protein synthesis by ribosomes and initiates the accumulation of hemoglobin and iron. Then through association with macrophages in erythroblastic islands, late stage erythroblasts enucleate and thereby become reticulocytes. At this stage, the cell should contain all the information necessary to complete the specialization into RBC and survive in the absence of transcription and new gene activity. A large pool of RNA is still present in reticulocytes, which mature within a few days by completing the synthesis of hemoglobin and other proteins that characterize mature RBCs and degrading internal organelles and “useless” proteins. In addition, reticulocytes migrate from the bone marrow into the blood circulation in which they terminate their maturation into RBCs. In this study, we extensively identified the proteins from the degradation machinery that actively eliminates superfluous species in reticulocytes and RBCs, i.e. all the subunits of the proteasome catalytic core (20 S proteasome) and regulatory complex (19 S proteasome) but also a wide range of E1, E2, and E3 enzymes from the ubiquitin-proteasome pathway (supplemental data 6). In particular, many E3 ubiquitin-protein ligases could be identified after sample treatment. These enzymes, which are responsible for targeting ubiquitination to specific substrate proteins and promoting their proteasomal degradation, should be particularly active in reticulocytes and RBCs where many proteins are no longer needed. It is questionable whether all the proteins that we identify in our final list are all genuine components of mature RBCs or whether some of them are incompletely degraded products that could be more specific to the reticulocytes also present in the sample. Although reticulocytes represent only 0.5–1% of the total cell population and should weakly contribute to the total protein content of the sample, minor proteins arising from these cells may be amplified because of the peptide libraries treatment. On the other hand, the degradation process initiated in reticulocytes is perpetuated in RBCs, and it is likely that a lot of proteins identified in this study are still present in the mature cells while being in the process of being degraded. It is also possible that many non-essential proteins are not necessarily degraded in the RBC. Indeed this would cost a lot of energy, whereas a better “solution” for the erythroblasts would be to just leave them intact but to specifically degrade some key proteins along their pathways. More focused studies would be needed to test this hypothesis and verify whether some of the identified proteins are present in the RBC in an active state or represent leftovers from the degradation process and to elucidate in more detail the mechanisms of this process. A lot of identified species, however, represent essential compo-

2 E. Boschetti and F. Berger, personal communication.
ponents of the mature cell, and some of them are discussed below.

Hemoglobins—Among the proteins of red blood cells, the most abundant is hemoglobin, which is in charge of oxygen transport from periphery to tissues. It represents 97–98% of the whole proteome. In vertebrates, hemoglobin is a heterotetramer composed of two types of subunits: the first type belongs to the $\alpha$-globin family ($\alpha$ or $\zeta$ chains) and is dependent on genes located on chromosome 16; the second type belongs to the $\beta$-globin family ($\beta$, $\gamma$, $\delta$, or $\varepsilon$ chains) with genes located on chromosome 11. Switches in globin gene expression appear during development; thus the embryonic $\zeta$ and $\varepsilon$ chains are replaced in the fetus by $\alpha$- and $\gamma$-globins as well as a little bit of $\delta$-globin. Finally in adult, $\alpha$ and $\beta$ chains are the most abundant, forming the HbA tetramer ($\alpha_2\beta_2$), but some $\gamma$- and $\delta$-globin chains are also present to form other types of hemoglobin. Mutations or deletion on globin gene loci can lead to different pathologies such as thalassemia (total or partial deficit of expression of one or several globin chains sometimes associated with abnormal persistence of fetal or embryonic globin forms) or other hemoglobinopathies (expression of globin variants) (25). In our study from healthy adult erythrocytes, the four chains $\alpha$, $\beta$, $\gamma$, and $\delta$ were identified in both non-treated and treated samples with a very high degree of coverage for the two major chains: $\alpha$, 99.3% and $\beta$, 97.3% (supplemental data 6). Interestingly the two embryonic chains, $\epsilon$ (in both samples) and $\zeta$ (in treated sample only), have also been identified confirming the observations made at the mRNA level that the corresponding genes are incompletely silenced during development (26, 27). In addition, in the present study we identified in the treated sample two other globin chains called $\mu$ and $\theta$. The gene of $\mu$ chain has been found recently on the $\alpha$-globin locus and has a high level of homology with the avian $\alpha^2$-globin (26). It has been detected in erythroid tissue by quantitative real time PCR (26) and recently identified by mass spectrometry by Pasini et al. (3), although its exact function is still unknown. Furthermore we identified, for the first time at the protein level, another globin chain called $\theta$. There has been some debate upon the existence or not of a protein product for the $\theta$ gene (28). However, the mRNA expression of this gene was demonstrated in fetal erythroid tissue (29), and it was shown to be a fetal/adult gene expressed like the $\alpha$-globin in fetus and adult but at a very low level (30). The identification in our study of these low expression globin chains confirms the ability of the peptide library treatment to reveal proteins usually hidden by major proteins ($\alpha$ and $\beta$ chains in this case).

Other Major RBC Pathways—The lifespan of RBCs is about 120 days during which the cell must generate energy to accomplish its vital functions. This is done by degrading glucose to lactate through the Embden-Meyerhof pathway with concomitant production of ATP (5). Efficient generation of ATP also relies on proper nucleotide metabolism. Indeed as RBCs cannot synthesize nucleotides de novo, they have evolved mechanisms that preserve the pool of adenine nucleotides in the cell by recycling adenine and adenosine into AMP and maintaining an equilibrium between AMP, ADP, and ATP. In addition, a small amount of glucose is also diverted and catabolized through the hexose monophosphate shunt (pentose phosphate pathway), which provides NADPH and reduced glutathione that are essential to protect the cell from oxidative damages. Protection against oxidation is critical in RBCs where the levels of oxygen and iron are very high. Therefore, any dysfunction in these pathways may have important deleterious effects, and many pathologies of the RBC are linked to inappropriate production of one enzyme involved in these metabolic functions. Those related to glycolysis or nucleotide metabolism hamper the generation of ATP and generally lead to chronic nonspherocytic hemolytic anemia, whereas those affecting the hexose monophosphate shunt will cause hemolytic episodes when RBCs are submitted to oxidative stress. In our study of the RBC cytoplasm, we identified most of the actors of these major pathways (supplemental data 6), particularly the enzymes involved in the most frequent metabolic RBC disorders such as pyruvate kinase, glucose-6-phosphate isomerase, or triose-phosphate isomerase. Although many of these proteins could be identified also in the non-treated sample, some particular species were detected only after peptide library treatment. This is the case for the PK-M form of pyruvate kinase. This enzyme catalyzes the second crucial ATP-generative step in glycolysis and exists as four isoenzymes (M1, M2, L, and R) encoded by two separate genes (PK-M and PK-LR), which are expressed in a tissue-specific manner (31, 32). The PK-LR gene codes for both PK-L (dominant form in liver) and PK-R (restricted to RBCs) through the use of alternate promoters. PK-R deficiency is the most common glycolytic enzyme defect associated with hereditary non-spherocytic hemolytic anemia. This gene product was identified in this study with high sequence coverage in the treated as well as non-treated RBC lysate. On the other hand, the PK-M gene codes for M1 and M2 isoenzymes by alternative splicing of the same RNA (33). The M1 protein predominates in skeletal muscle, heart, and brain, whereas the M2 isoenzyme is found primarily in rapidly proliferating fetal tissues. The M2 form is also found in erythroblasts where it is progressively replaced by the R form as the cells mature into RBCs (34). In pathologies related to PK deficiency, persistence of the fetal M2 form can be observed in mature cells possibly to compensate for the lack of the R isozyme. Moreover PK-M2 has been shown to be overexpressed in tumor cells (35) and is evaluated as a diagnostic biomarker in many cancer types (36–40). The fact that we could specifically detect this form in the treated lysate from RBC, although it is probably present at very low amounts in these healthy, mature cells, illustrates the performance of the method for low abundance species identification and its potential for biomarker discovery.
Conclusion—By using combinatorial peptide libraries and powerful nano-LC-MS technology, we could map extensively the cytosolic RBC proteome and demonstrate the presence in these cells of a large body of previously undetected species. Given the high number of proteins identified and the wide range of physiological functions to which they are related, more focused studies would be needed to understand better whether these proteins still play a role in the mature cell. Clearly some minor proteins reported here are inherited from the erythroblast, like for example the BCI-XL anti apoptotic molecule, which is essential for survival of precursor cells from the erythroid lineage during the differentiation process (41), or the Emp protein, recently shown to be important for nuclear extrusion from erythroblasts and their terminal maturation (42). However, even if these proteins are no longer essential in the mature cell, their identification remains of interest. It shows that proteasomal degradation of many species is not complete or maybe not even performed at all and that the RBC cytosol can constitute an easily accessible reservoir of biomarkers carrying potential information about dysfunctions taking place earlier in the differentiation process. This can be interesting in particular to study myeloproliferative disorders characterized by abnormal and malignant transformations of the precursor hematopoietic cells as in the case for example of erythroleukemia, erythrocytosis, or polycythemia vera, pathologies associated with increased hemapoiesis and overproduction of RBCs (43). In the case of polycythemia vera, a somatic mutation of the JAK2 gene leading to dysfunction of the corresponding kinase has been shown to be involved in pathogenesis (44). However, no convenient serum biomarker exists today for this disease, and screening the RBC content by differential proteomics could help in finding more efficient diagnosis tests. The combinatorial peptide library approach could be a powerful tool for such biomarker research projects as it significantly increases the number of detectable species in global protein profiling strategies. In addition, we hope that the important data set of proteins provided here will prove to be useful to the community for more dedicated studies focusing on specific protein species and aiming at understanding better the molecular bases of RBC biology, terminal erythroid differentiation, hemoglobin switching, or RBC diseases. Indeed extensive proteomics data sets can become the basis of targeted mass spectrometry analyses in which the instrument is used to monitor and quantify with high sensitivity and specificity particular proteotypic peptides uniquely associated with a protein of interest (45). The mass spectrometer can indeed be operated in the multiple reaction monitoring mode where the monitoring of “transitions” composed of one parent ion and one fragment ion allows the very sensitive detection of the selected proteins in a complex mixture. This emerging targeted work flow thus relies on proper selection of an observable proteotypic peptide and of an optimal transition, which can be more efficient if based on experimental data. To this end, we made acces-
sible all the MS and MS/MS data acquired during this study. The final protein lists, peptide information, and MS/MS spectral data can thus be browsed through the MFPaQ interface at http://proteomique.ipbs.fr/rbc.

Acknowledgments—We are grateful to Dr. François Berger and Dr. Pascal Mossuz for fruitful discussions.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

§§ Supported by the CNRS, the Génopole Toulouse Midi-Pyrénées, the Région Midi-Pyrénées, and grants from the Institut National du Cancer, Agence Nationale de la Recherche, and Fondation pour la Recherche Médicale. To whom correspondence may be addressed: Inst. de Pharmacologie et de Biologie Structurale, 205 route de Narbonne, 31077 Toulouse cedex 4, France. E-mail: bernard.monsarrat@ipbs.fr.

REFERENCES

1. Delaunay, J. (2007) The molecular basis of hereditary red cell membrane disorders. Blood Rev. 21, 1–20

2. Walensky, L. D., Narla, M., and Lux, S. E., IV (2003) Disorders of the red blood cell membrane, in Blood: Principles and Practice of Hematology (Handin, R. I., Lux, S. E., and Stossel, T. P., eds) pp. 1709–1858, Lippincott Williams & Wilkins, Philadelphia

3. Pasini, E. M., Kirkaegda, M., Mortensen, P., Lutz, H. U., Thomas, A. W., and Mann, M. (2006) In-depth analysis of the membrane and cytosolic proteome of red blood cells. Blood. 108, 791–801

4. van Wijk, R., and van Solinge, W. W. (2005) The energy-less red blood cell is lost: erythrocyte enzyme abnormalities of glycolysis. Blood. 106, 4034–4042

5. Eber, S. W. (2003) Disorders of erythrocytes glycolysis and nucleotide metabolism, in Blood: Principles and Practice of Hematology (Handin, R. I., Lux, S. E., and Stossel, T. P., eds) pp. 1887–1920, Lippincott Williams & Wilkins, Philadelphia

6. Vulimaty, T. J., and Luzzatto, L. (2003) Glucose-6-phosphate dehydrogenase deficiency and related disorders, in Blood: Principles and Practice of Hematology (Handin, R. I., Lux, S. E., and Stossel, T. P., eds) pp. 1921–1950, Lippincott Williams & Wilkins, Philadelphia

7. Kaknishaivili, D. G., Bulla, L. A., Jr., and Goodman, S. R. (2004) The human erythrocyte proteome: analysis by ion trap mass spectrometry. Mol. Cell. Proteomics 3, 501–509

8. Bhattacharya, D., Mukhopadhyay, D., and Chakrabarti, A. (2007) Hemoglobin depletion from red blood cell cytosol reveals new proteins in 2-D gel-based proteomics study. Proteomics Clin. Appl. 1, 561–564

9. Prabakaran, S., Wengenroth, M., Lockstone, H. E., Lilley, K., Leweke, F. M., and Bahn, S. (2007) 2-D DIGE analysis of liver and red blood cells provides further evidence for oxidative stress in schizophrenia. J. Proteome Res. 6, 141–149

10. Thulasiramam, V., Lin, S., Gheorghiu, L., Lathrop, J., Lomas, L., Hammond, D., and Boschetti, E. (2005) Reduction of the concentration difference of proteins in biological liquids using a library of combinatorial ligands. Electrophoresis 26, 3561–3571

11. Guerrier, L., Thulasiramam, V., Castagna, A., Fortis, F., Lin, S., Lomas, L., Righetti, P. G., and Boschetti, E. (2006) Reducing protein concentration range of biological samples using solid-phase ligand libraries. J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 833, 33–40

12. Castagna, A., Cecconi, D., Sennels, L., Rappapalter, J., Guerrier, L., Fortis, F., Boschetti, E., Lomas, L., and Righetti, P. G. (2005) Exploring the
hidden human urinary proteome via ligand library beads. J. Proteome Res. 4, 1917–1930

13. Sennels, L., Salek, M., Lomas, L., Boschetti, E., Righet, P. G., and Rapp-silber, J. (2007) Proteomic analysis of human blood serum using peptide library beads. J. Proteome Res. 6, 4055–4062

14. Guerrier, L., Claverol, S., Fortis, F., Rinalducci, S., Tempier, A. M., Antonioli, P., Jandrot-Perrus, M., Boschetti, E., and Righet, P. G. (2007) Exploring the platelet proteome via combinatorial, hexapeptide ligand libraries. J. Proteome Res. 6, 4290–4303

15. Fortis, F., Guerrier, L., Areces, L. B., Antonioli, P., Hayes, T., Carrick, K., Hammond, D., Boschetti, E., and Righet, P. G. (2006) A new approach for the detection and identification of protein impurities using combinatorial solid phase ligand libraries. J. Proteome Res. 5, 2577–2585

16. Antonioli, P., Fortis, F., Guerrier, L., Rinalducci, S., Zolla, L., Righet, P. G., and Boschetti, E. (2007) Capturing and amplifying impurities from purified recombinant monoclonal antibodies via peptide library beads: a proteomic study. Proteomics 7, 1624–1633

17. Righetti, P. G., Boschetti, E., Lomas, L., and Citterio, A. (2006) Protein Equalizer Technology: the quest for a “democratic proteome”. Proteomics 6, 3980–3992

18. Righetti, P. G., and Boschetti, E. (2007) Sherlock Holmes and the proteome—a detective story. FEBS J. 274, 897–905

19. Bouyssie, D., de Peredo, A. G., Mouton, E., Albigot, R., Roussel, L., Ortega, N., Cavour, C., Burel-Schiltz, O., Girard, J. P., and Monsarrat, B. (2007) Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells. Mol. Cell. Proteomics 6, 1621–1637

20. Durr, E., Yu, J., Krasinska, K. M., Carver, L. A., Yates, J. R., Testa, J. E., Oh, P., and Schnitzer, J. E. (2004) Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture. Nat. Biotechnol. 22, 985–992

21. Liu, H., Sadygov, R. G., and Yates, J. R., III (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal. Chem. 76, 4193–4201

22. Shen, Y., Kim, J., Strittmatter, E. F., Jacobs, J. M., Camp, D. G., II, Fang, R., Tolle, N., Moore, R. J., and Smith, R. D. (2005) Characterization of the human blood plasma proteome. Proteomics 5, 4034–4045

23. de Godoy, L. M., Olsen, J. V., de Souza, G. A., Li, G., Mortensen, P., and Mann, M. (2006) Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. Genome Biol. 7, R50

24. Guerrier, L., Claverol, S., Finzi, L., Paye, F., Fortis, F., Boschetti, E., and Houssset, C. (2007) Contribution of solid-phase hexapeptide ligand libraries to the repertoire of human bile proteins. J. Chromatogr. A 1176, 192–205

25. Forget, G. F., and Olivier, N. F. (2003) Hemoglobin synthesis and the thalassemias, in Blood: Principles and Practice of Hematology (Handin, R. I., Lux, S. E., and Stossel, T. P., eds) pp. 1503–1596, Lippincott Williams & Wilkins, Philadelphia

26. Goh, S. H., Lee, Y. T., Bhanu, N. V., Cam, M. C., Desper, R., Martin, B. M., Moharram, R., Gheiner, R., and Miller, J. L. (2005) A newly discovered human α-globin gene. Blood 106, 1466–1472

27. Albitar, M., Peschle, C., and Liebhaber, S. A. (1989) η, ζ, and ε globin messenger RNAs are expressed in adults. Blood 74, 629–637

28. Clegg, J. B. (1987) Can the product of the η gene be a real globin? Nature 329, 465–466

29. Leung, S., Proudfoot, N. J., and Whelaw, E. (1987) The gene for η-globin is transcribed in human fetal erythroid tissues. Nature 329, 551–554

30. Albitar, M., Care, A., Peschle, C., and Liebhaber, S. A. (1992) Developmental switching of messenger RNA expression from the human α-globin cluster: fetal/adult pattern of η-globin gene expression. Blood 80, 1586–1591

31. Zanella, A., Ferro, E., Bianchi, P., and Valentini, G. (2005) Red cell pyruvate kinase deficiency: molecular and clinical aspects. Br. J. Haematol. 130, 11–25

32. Zanella, A., Bianchi, P., and Ferro, E. (2007) Pyruvate kinase deficiency. Haematologica 92, 721–723

33. Naguchi, T., Inoue, H., and Tanaka, T. (1986) The M1- and M2-type isozymes of rat pyruvate kinase are produced from the same gene by alternative RNA splicing. J. Biol. Chem. 261, 13807–13812

34. Takegawa, S., and Miwa, S. (1984) Change of pyruvate kinase (PK) isozymes in classical type PK deficiency and other PK deficiency cases during red cell maturation. Am. J. Hematol. 16, 53–58

35. Mazurek, S., Grimm, H., Boschek, C. B., Vaupel, P., and Eigenbrodt, E. (2002) Pyruvate kinase type M2: a crossroad in the tumor metabolome. Br. J. Nutr. 87, Suppl. 1, S23–S29

36. Schneider, J. (2006) Tumor markers in detection of lung cancer. Adv. Clin. Chem. 42, 1–41

37. Ahmed, A. S., Dew, T., Lawton, F. G., Papadopoulos, A. J., Devaja, O., Raju, K. S., and Sherwood, R. A. (2007) M2-PK as a novel marker in ovarian cancer. A prospective cohort study. Eur. J. Gynaecol. Oncol. 28, 23–28

38. Hardt, P. D., Mazurek, S., Toeppler, M., Schlierbach, P., Bretzel, R. G., Eigenbrodt, E., and Kloer, H. U. (2004) Faecal tumour pyruvate kinase: a new, sensitive screening tool for colorectal cancer. Br. J. Cancer 91, 980–984

39. Ugurel, S., Bell, N., Sucker, A., Zimpfer, A., Rittgen, W., and Schadendorf, D. (2005) Tumor type M2 pyruvate kinase (TuM2-PK) as a novel plasma tumor marker in melanoma. Int. J. Cancer 117, 825–830

40. Kumar, I., Tapuria, N., Kirmani, N., and Davidson, B. R. (2007) Tumour M2-pyruvate kinase: a gastrointestinal cancer marker. Eur. J. Gastroent. Hepatol. 19, 265–276

41. Opferman, J. T. (2007) Life and death during hematopoietic differentiation. Curr. Opin. Immunol. 19, 497–502

42. Soni, S., Bala, S., Gwynn, B., Sahr, K. E., Peters, L. L., and Hanspal, M. (2006) Absence of erythroblast macromage protein (Emp) leads to failure of erythroblast nuclear extrusion. J. Biol. Chem. 281, 20181–20189

43. Chen, G., and Prchal, J. T. (2006) Polycythemia vera and its molecular basis: an update. Best Pract. Res. Clin. Haematol. 19, 387–397

44. James, C., Ugo, V., Le Couedic, J. P., Staerk, J., Delhommeau, F., Lacout, C., Garcon, L., Raslova, H., Berger, R., Benceceur-Griscelli, A., Villeval, J. L., Constantinescu, S. N., Casadevall, N., and Vainchenker, W. (2005) A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. Nature 434, 1144–1148

45. Malmstrom, J., Lee, H., and Aebersold, R. (2007) Advances in proteomic workflows for systems biology. Curr. Opin. Biotechnol. 18, 378–384