UMI-linked consensus sequencing enables phylogenetic analysis of directed evolution

Paul Jannis Zurek, Philipp Knyphausen, Katharina Neufeld, Ahir Pushpanath & Florian Hollfelder

The success of protein evolution campaigns is strongly dependent on the sequence context in which mutations are introduced, stemming from pervasive non-additive interactions between a protein’s amino acids ('intra-gene epistasis'). Our limited understanding of such epistasis hinders the correct prediction of the functional contributions and adaptive potential of mutations. Here we present a straightforward unique molecular identifier (UMI)-linked consensus sequencing workflow (UMIC-seq) that simplifies mapping of evolutionary trajectories based on full-length sequences. Attaching UMIs to gene variants allows accurate consensus generation for closely related genes with nanopore sequencing. We exemplify the utility of this approach by reconstructing the artificial phylogeny emerging in three rounds of directed evolution of an amine dehydrogenase biocatalyst via ultrahigh throughput droplet screening. Uniquely, we are able to identify lineages and their founding variant, as well as non-additive interactions between mutations within a full gene showing sign epistasis. Access to deep and accurate long reads will facilitate prediction of key beneficial mutations and adaptive potential based on in silico analysis of large sequence datasets.
Directed evolution has emerged as a viable route to improve biocatalysts, as recognized by the Nobel Prize in Chemistry in 2018\(^2\), but governing the intrinsically random process of an evolution campaign to ensure its success is a daunting challenge. Evolution can be conceptualized as an exploration of “sequence space” along possible trajectories that lead to useful sequences with desirable properties\(^2,3\). The combination of ultrahigh throughput screening and next-generation sequencing is a powerful way of mapping such combinatorial explorations of protein function. The knowledge obtained may be used to steer experimentation into promising regions within the “vastness of sequence space”, simultaneously helping to understand fundamental concepts underlying sequence-function relationships\(^4-6\). Examples of this approach have revealed e.g. that epistasis is responsible for shaping the local fitness landscape of GFP\(^7\), provided insight into the trade-off between enzyme solubility and activity\(^8\) or enabled the structural prediction of different small protein binders\(^9,10\). Similarly, data parameterization starting from sequence information enabled the design of selective peptide blockers using machine-learning approaches\(^11\). There are indications that gathering extensive sequence datasets during evolutionary transitions is crucial for the development of machine-learning algorithms to enable data-driven predictions for protein engineering based on statistical models of protein function\(^12-16\).

Evolutionary lineages are characterized by non-linear interactions between mutations within one gene that can only be inferred from accurate full-length sequence analysis of gene variants. Such “epistatic” interactions are key determinants of protein evolution and synergy or incompatibility of mutations determines which pathways are accessible in sequence space\(^17-19\). To delineate and understand the principles of these interactions, high throughput and accurate full-length sequences of gene variants are necessary. While Sanger sequencing can provide accurate sequence information with read lengths covering the extent of typical genes (0.9–1.4 kb)\(^20\), it is very cost-intensive when high throughput is required. Sequencing-by-synthesis platforms such as Illumina, on the other hand, offer great accuracy and throughput, but their primary readout is limited to short reads (up to 300 bp), falling far short of the average gene length. Illumina sequencing is thus unable to directly capture simultaneously occurring mutations across a full gene. This dilemma can be overcome with synthetic long reads, which are generated by associating multiple short reads into long sequences. However, synthetic long-read methods for amplicon sequencing are known to suffer from uneven coverage\(^21-23\) or high chimera rates\(^24,25\) resulting in a substantially reduced number of correct full-length sequences and thus effectively lowered throughput (for more details see “Discussion”). Third-generation single-molecule sequencing technologies (such as those developed by Pacific Biosciences and Oxford Nanopore Technologies) are able to generate truly long reads at high throughput but suffer from high error rates, preventing the accurate assignment of single point mutations to individual reads (e.g. for nanopore sequencing with an accuracy of 85–95% at least 50 in 1000 bases are assigned incorrectly)\(^26\). The extensive read lengths enable direct association of co-occurring mutations; however, generation of consensus sequences from multiple reads is necessary to increase accuracy.

Accurate consensus sequences are commonly generated for genome assemblies, where every sequencing read represents a unique fragment that overlaps with many others, facilitating stacking for accurate consensus generation\(^27,28\). By contrast, libraries in directed protein evolution comprise rather closely related members. Here the sequence diversity in randomized gene libraries is typically low: protein variants with a few or even a single point mutation can be functionally different. An erroneous raw third-generation sequencing read cannot reliably be assigned to its template, when multiple template candidates differing only in few point mutations are possible. This inability to assign reads to any one parental molecule therefore renders consensus-based “polishing” from unlinked reads impossible. Similar samples in which individual members differ merely in a few nucleotides can also be found in e.g. immune repertoires, metagenomic 16S amplicons or medical diagnostics, calling for a suitably high-quality sequencing solution. Current approaches to increase the accuracy of nanopore sequencing rely on the generation of concatemerized sequencing templates with multiple copies of the same gene variant forming a single DNA molecule. This physical copy-linkage results in long reads containing multiple “sub-reads” of the same gene variant, based on which a consensus sequence can be generated with up to 99.5% accuracy\(^29\) (i.e. 5 in 1000 bases are assigned incorrectly), depending on gene and concatemer length. However, this accuracy is still insufficient to distinguish point mutations and practically the workflow includes multiple delicate steps that require careful handling of high-molecular weight DNA molecules.

To achieve higher nanopore sequencing accuracy, we developed a workflow generating a high-quality consensus based on a bioinformatically traceable sequence link between descendant reads and their molecule of origin by introducing random DNA barcodes (i.e. unique molecular identifiers, UMIs) prior to amplification for sequencing. UMIs are widely used to retain information on true molecule numbers before PCR amplification. Specifically, the distinct template identity marked by the UMI enables confident identification and quantification of PCR duplicates\(^30\). This method has been applied e.g. to achieve unbiased counting of molecules for expression data in RNA-seq experiments\(^31,32\) and the accurate quantification of rare mutations\(^33,34\). Here, we leverage UMI-tags to assign erroneous nanopore reads to their molecule of origin, facilitating clustering for accurate consensus formation even when starting with a pool of highly similar sequences (e.g. a library of gene variants in protein evolution generated by error-prone PCR). Such sequences typically differ in only a few point mutations and can currently not be distinguished reliably in an ordinary nanopore sequencing output. We apply our workflow to protein engineering and demonstrate the analysis of high-quality full sequence outputs through rounds of ultrahigh throughput directed evolution of an amine dehydrogenase (AmDH), tracking the emerging phylogeny—or the “walk through sequence space”\(^2\) —towards higher activity in directed protein evolution.

Results

UMIC-seq is a nanopore sequencing workflow for libraries with few nucleotide variations. To achieve easy control over molecule count and consensus coverage every library member was tagged with a unique molecular identifier (UMI) (Fig. 1). The UMIs allow individual sequencing reads to be assigned to their template molecule, enabling the generation of UMI-linked consensus sequences with very high final accuracy (>99.99%). Specifically, we increased sequencing accuracy by tagging and selectively amplifying single molecules with UMIs, which enable the generation of consensus sequences from many reads per variant. The UMIs have to be highly diverse to allow faithful assignment of raw, low-accuracy nanopore reads to the corresponding variant. We have performed simulations of clustering efficiencies using our UMIC-seq workflow under different error-rate regimes and library sizes to estimate suitable UMI lengths for our application (Supplementary Fig. 1). Based on these simulations, fully randomized 50 bp sequences were chosen as UMIs for subsequent experiments, as they were predicted to enable
complete recollection of clusters over a broad range of error rates (Supplementary Fig. 1a) as well as homogeneous clusters in libraries of up to a million variants (Supplementary Fig. 1b). The UMIs were incorporated using primers in two PCR cycles, with deliberately low amplification so that potential PCR bias was reduced compared to other methods.

As the number of theoretically possible UMI sequences \((4^{50})\) far outweighs the number of template molecules in the PCR reaction, each template molecule will be uniquely tagged. The number of UMI-tagged molecules generated, however, is too large to achieve deep sequencing coverage. It is thus necessary to selectively amplify a small subset of UMI-variant combinations. Uniquely, the UMI-variant complexity is reduced by a transformation step into cells, because an exact number of molecules is represented in the number of transformant colonies, directly measurable by a transformation step into cells, because an exact number of molecules is represented in the number of transformant colonies, directly measurable.

With this filtering threshold, and using 180 pairs of Sanger reads (a total of 173 kb) as reference, we calculate a mean per-base error rate of 0.008% from as little as 35-fold sequencing coverage (a total of 173 kb) as reference, we calculate a mean per-base error rate of 0.008% from as little as 35-fold sequencing coverage (Supplementary Fig. 2b). We identified 654 true mutations via Sanger sequencing of the 180 reference variants, of which 98% were correctly identified in the UMIC-seq workflow. The mean error rate per mutation (or single-nucleotide variation) for the nanopore consensus sequences were 1.99% false negatives and 0.16% false positives. The relatively higher number of false negatives is likely due to errors in clustering, resulting in mixed clusters for which no mutations are called (Supplementary Note 1). The resulting slight inflation of WT-like non-mutated sequences in the dataset does not interfere with the analysis of phylogeny and trajectory in protein evolution.

**Fig. 1 Sequencing strategy for amplicon libraries with UMI-linked nanopore long reads (UMIC-seq).** An amplicon library with similar members, such as a plasmid pool of improved enzyme variants, an immune repertoire or metagenomic 16S sequences (1) is used as input to UMI-tagging and barcoding by PCR (2). The resulting products have unique sequence identifiers (UMI) for clustering and barcodes for multiplexing, as well as homology overhangs used for Gibson assembly. Gibson assembly provides plasmids for transformation, which enables the selection of a subset of the library, obtaining selectively amplified UMI-variant combinations (3). This sample is then subjected to nanopore sequencing (4) and data processing via barcode- and UMI-based clustering followed by variant calling, yielding the desired variant sequences with high accuracy (5).
Accurate long-read sequencing applied to the microfluidic evolution of an amine dehydrogenase (AmDH) biocatalyst. To exemplify the utility of the UMIC-seq workflow, we applied directed evolution to an AmDH (Fig. 2), a valuable biocatalyst for the synthesis of chiral amines. This enzyme can be screened in an ultrahigh-throughput droplet microfluidic assay (>10^5 droplets per hour), which leads to a large harvested hit diversity exceeding the limits of conventional hit analysis based on Sanger sequencing. Our objective was to track the accumulation of mutations in pools of variants and retain information on the co-occurrence of mutations over three rounds of directed evolution (Supplementary Fig. 3). Specifically, 1-3 mutations were introduced per gene and per round of directed evolution using error-prone PCR of the parental AmDH sequence (in the first round) and the obtained hit pools (for the consecutive rounds). The respective enzymes were expressed in *E. coli* that were compartmentalized in droplets as single cells, where lysis agent liberated the expressed enzyme and encounter with the reaction substrate allowed the enzymatic reaction to proceed. Product formation could then be detected as an optical read-out (Fig. 2). In each round of screening, the 1000 best hits out of ~250,000 variants were isolated via absorbance-activated droplet sorting (AADS, Supplementary Fig. 4) and sequenced using the UMI-linked long-read workflow with an Oxford Nanopore MinION flow cell.

**Fig. 2 Ultrahigh-throughput droplet microfluidic evolution of an amine dehydrogenase (AmDH) biocatalyst.**

**a** Coupled reaction to measure AmDH activity in droplets. The enzyme AmDH catalyzes the deamination of (R)-1-methyl-3-phenylpropylamine to 4-phenyl-2-butane, thereby reducing NAD\(^+\) to NADH and releasing NH\(_3\). NADH is regenerated upon reduction of WST-1, mediated by 1-methoxy-5-methylphenazinium methyl sulfate (mPMS). Reduced WST-1 exhibits absorbance at 455 nm, which is detectable by absorbance-activated droplet sorting (AADS) on chip. **b** A cycle of directed evolution using error-prone PCR was followed by transformation and compartmentalization of cells in droplets. Cell lysis upon droplet formation liberates AmDH and reaction progress was determined based on the formation of colored downstream product in a coupled reaction. The best ~0.5% of variants in the library were selected by AADS and the respective plasmid pool was isolated.

Long-read sequencing information uniquely informs protein engineering. A hot-spot analysis based on the frequency of mutation during the experimental directed evolution not only revealed several positions randomly scattered over the full length of the protein, but also indicated accumulation close to the loop covering the substrate binding pocket (Fig. 4a). This underlines the importance of less obvious distal mutations to protein function. Hot-spot analysis based on evolutionary conservation from established sequence data is often used as guidance for site-directed protein engineering; however, a hot-spot analysis does not report on the compatibility of the hot-spots with each other and with mutations in the rest of the protein. In contrast to hot-spot analysis, the sequence information obtained in this workflow describes multiple mutations in each variant over the full length of the gene (Supplementary Fig. 3): now improved variants with multiple mutations can be identified. Consequently, epistatic interactions, which are prevalent in protein evolution and difficult or impossible to predict, can be detected and the...
Distributed Stochastic Neighbour Embedding, tSNE) of variants identified in droplet screening illustrates possible trajectories in a sequence similarity network. Spots represent all unique variants identified within the pools of hits. Distinct clusters emerge over the course of evolution, as indicated by the color code showing the round of first emergence of each sequence. Size of wild type increased for visualization.

Included in the core sets of mutations that they share with the founder variant, so that these mutations are present in all variants within the cluster. Spot sizes correspond to total sequence count of the respective variant. The clusters are defined by multiple mutations and thus can only be inferred based on long sequence reads that inform phylogeny of possible trajectories during experimental directed evolution.

Fig. 3 Phylogenetic analysis of directed evolution in sequence similarity networks. a Full graph color coded by the round of evolution in which mutant sequences were first recorded. A graphical representation (with t-Distributed Stochastic Neighbour Embedding, tSNE) of variants identified in droplet screening illustrates possible trajectories in a fitness landscape: spots represent all unique variants identified within the pools of hits. Distinct clusters emerge over the course of evolution, as indicated by the color code showing the round of first emergence of each sequence. Size of wild type increased for visualization. b Analysis of founder variants. Part of the sequence similarity network is shown with the color indicating the corresponding core set of mutations that they share with the ‘founder variant’, so that these mutations are present in all variants within the cluster. Spot sizes correspond to total sequence count of the respective variant. The clusters are defined by multiple mutations and thus can only be inferred based on long sequence reads that inform phylogeny of possible trajectories during experimental directed evolution.

Discussion

Interestingly, an example of sign epistasis can be seen in the founder variants A64E R102S E323V. The mutation E323V individually decreases lysate deamination activity to 15.7 ± 1.1% (95% CI) of the non-mutated parent, but it has a beneficial impact in the background of the founder variant A64E R102S, increasing its lysate deamination activity to 154 ± 33% (95% CI) of the A64S R102S variant (Fig. 4b). The basis of the interaction of these three mutations is not obvious, as they are far apart (Fig. 4a highlights the distance of at least 18 Å between the mutations that are shown to interact epistatically). The identification of such functionally relevant long-range interactions is only possible when long sequence reads are available: in our case the epistatic interaction extends across 780 bp (or 260 amino acids) in the sequence. This distance exceeds the conventional length limitations of second-generation sequencing that can only be overcome by assembling synthetic long reads in silico (see “Discussion”). To the best of our knowledge, the relationship displayed in Fig. 4 is the longest-ranging epistatic interaction in a protein identified in a high throughput mapping experiment to date. Conventional hot-spot analysis based on short-read sequencing (i.e. analyses such as the study of glycosidase catalysis via microfluidic deep mutational scanning)\(^{44}\), might have highlighted important residues but
Fig. 4 Analysis of epistasis in variant A64E R102S E323V. a Hot-spot analysis. Positional enrichments mapped onto the structure of a homologous AmDH (PDB ID: 1C1D). If the frequency of mutation at a certain position increased over the rounds of directed evolution, its enrichment factor is color coded on the structure from green (low) to yellow to red (high). Enrichment was calculated from round 1 to round 2 for variants that persist into round 3. Positions at which mutations are enriched can be identified over the full range of the enzyme, with many enriched positions e.g. in the loop covering the substrate binding pocket. Positions exhibiting epistatic interaction (variant A64E R102S E323V) are highlighted. b Sign epistasis in variant A64E R102S E323V detected by long-read sequencing. Graph showing lysate deamination activities (initial rates) of constituent variants relative to the non-mutated lysate assay.

Characterization of the founder variant A64E R102S E323V. The particularly successful founder variant A64E R102S E323V and the contributing mutation combinations were characterized with regard to their catalytic properties and performance as biocatalysts (Table 1). Interestingly, the mutation A64E has a strong contribution to an altered catalytic profile, increasing both the $K_M$ and the $k_{cat}$. The $k_{cat}$ of variant A64E is almost 4-fold increased, indicating a strong selection for fast turnover at high substrate concentrations. This is in accordance with the directed evolution principle "you get what you screen for"45, since the absorbance detection in droplets requires high substrate turnover. Thus, the droplet screening assay is performed with long incubation times at high substrate concentration, putting selection pressure on high $k_{cat}$ values. Additionally, mutation A64E increases expression strength which is an important factor in lysate assays, such as the droplet microfluidic screening, as it increases the amount of available protein in the compartment. Mutation E323V, on the other hand, is deleterious in the lysate assay (Fig. 4b), stemming from a reduced catalytic efficiency as well as a stark reduction in expression (Table 1). The decrease in protein expression caused by mutation E323V, however, is compensated by mutation A64E, laying the foundation of the epistatic interactions. When E323V is combined with the other mutations, a decreased $K_M$ value results in higher catalytic efficiencies compared to mutation A64E alone, further increasing the overall performance of the combined variant in the lysate activity assay.

This increased performance is confirmed in a biocatalytic assay, the determination of the enzyme’s total turnover number. The total turnover number was determined in the industrially relevant amination reaction direction at high substrate concentration. Here, the variant A64E R102S E323V again outperforms not only the wild type but also mutation A64E by a wide margin, indicating the suitability of the proposed approach to protein engineering to solve an applied biocatalytic challenge.

While the strongly activating mutation A64E would have been found by summing up mutation frequencies in short reads, the

### Table 1 Characterization of A64E R102S E323V core set of mutations.

| Variant          | Steady state kinetics | $T_m$ (°C) | Soluble expression (%) | Total turnover number |
|------------------|-----------------------|------------|------------------------|-----------------------|
|                  | $k_{cat}$ (s$^{-1}$) | $K_M$ (mM) | $k_{cat}/K_M$ (s$^{-1}$ mM$^{-1}$) |                        |
| WT               | 1.55 ± 0.02           | 0.40 ± 0.02 | 3.84 ± 0.06           | 50.5 ± 0.17           | 97 ± 0.25            | 20,100 ± 1,180       |
| A64E             | 6.12 ± 0.05           | 2.17 ± 0.06 | 2.83 ± 0.03           | 50.4 ± 0.08           | 99 ± 0.40            | 57,400 ± 730         |
| R102S            | 1.72 ± 0.03           | 0.33 ± 0.03 | 5.27 ± 0.08           | 49.7 ± 0.09           | 96 ± 0.11            | 22,000 ± 330         |
| E323V            | 1.20 ± 0.01           | 0.37 ± 0.02 | 3.28 ± 0.05           | 51.0 ± 0.09           | 79 ± 3.34            | 16,400 ± 700         |
| A64E R102S       | 5.38 ± 0.07           | 1.86 ± 0.07 | 2.90 ± 0.04           | 48.1 ± 0.08           | 97 ± 0.22            | 51,500 ± 720         |
| A64E E323V       | 4.94 ± 0.09           | 1.62 ± 0.09 | 3.05 ± 0.06           | 49.9 ± 0.05           | 94 ± 0.47            | 66,300 ± 2,990       |
| R102S E323V      | 1.36 ± 0.02           | 0.38 ± 0.02 | 3.62 ± 0.05           | 49.9 ± 0.07           | 98 ± 0.18            | 16,200 ± 2,030       |
| A64E R102S E323V | 4.23 ± 0.07           | 1.27 ± 0.07 | 3.34 ± 0.06           | 48.9 ± 0.09           | 98 ± 1.45            | 65,900 ± 380         |

Source data are provided as a Source Data file. All values are given with standard deviations from three independent technical replicates.

*a Steady state kinetics: Buffer: 100 mM Glycine-KOH pH 10. Co-substrate: 2.5 mM NAD$^+$. Temperature: 22 °C. Substrate: R-1-methyl-3-phenylpropylamine 0 mM to 12.8 mM. Supplementary Fig. 6.

*b Melting temperature: Differential scanning fluorimetry with SYPRO orange. Supplementary Fig. 7.

*c Soluble expression: Protein expression performed at 20°C for 16 h. Soluble fraction compared to pellet via gel densitometry.

*d Total turnover number: Small scale transformation with 50 mM 4-phenyl-2-butanone and 0.1 mM enzyme. Reaction performed at 30 °C for 72 h with glucose dehydrogenase cofactor recycling.
compatibility and the enhancing effects of other residues would have been missed in a hot-spot analysis. Consequently, third-generation reads of entire genes (typically 0.9–1.4 kb) are necessary in order to discover complete sets of epistatic interactions in proteins. Enabling such long reads opens the door to resolving complex interaction networks that will help to predict protein structures, train machine-learning algorithms for protein engineering or study the fundamental constraints of protein evolution. The comprehensive investigation of such phenomena by the simple nanopore sequencing workflow of Fig. 1 will allow the discovery of patterns typical for short- and long-range epistasis, providing a valuable thesaurus of information for future protein engineering.

Discussion

The sequencing workflow presented here provides highly accurate long-read sequences suitable for the analysis of entire protein engineering datasets with single-base precision at unprecedented low cost. With a MinION flow cell capacity of 10 Gb and assuming a gene of interest of 2 kb length, our workflow can generate approximately 100,000 accurate consensus sequences at 50-fold coverage. This entails a cost per mutant sequence of <1,4, a 450-fold reduction compared to around 5 USD using classical Sanger sequencing (Supplementary Note 2), with virtually no limit to gene lengths that can be analyzed.

Using regular Illumina short reads, genes have to be split up into fragments and association between random mutations over a full gene length are lost. The longest possible DNA stretch that can be sequenced straightforwardly with second-generation technology is 600 bp, obtained by associating forward and reverse sequencing reads. To overcome these limitations, synthetic long reads are constructed from short reads by linking them via compartmentalized barcoding strategies, which has been commercialized e.g. by Moleculo or 10x Genomics. However, while these methods are suitable for genomics, they do not meet the challenges of amplicon sequencing. In plate-based methods fragments are pooled, making it unsuitable for libraries of highly similar variants, whereas microfluidic droplet-compartmentalized barcoding would require a reduction in fragment input concentration to maintain the monoclarity of droplet compartments (i.e. reduce collision of multiple variants per droplet), thus limiting the throughput of the method. Current approaches designed for one-pot synthetic long-read amplicon sequencing are limited in maximal fragment length by the short-read bridge amplification requirements, have high numbers of chimeric reads or a low percentage of full length. Uneven coverage over the full length of the parent fragment results in these methods needing high sequencing coverage and still often failing at generating full-length consensus sequences.

These issues are reflected in current investigations of pairwise epistatic relationships in proteins. These studies focus, in most cases, on short domains (e.g. 56 amino acids in the IgG-binding domain of protein G, 37 amino acids in the WW domain of Yap, 32 amino acids at the interface of Fos and Jun or 25 amino acid blocks in the second RRM domain of Pab). Capturing the entire protein sequence would be necessary to reveal the complete, presumably more complex epistatic interactions in entire proteins rather than domains. The longest stretch of sequence analyzed for multiple simultaneous mutations in a combinatorial experiment are the 238 amino acids of GFP, that were analyzed to describe its fitness landscape. Here, shorter primary reads were barcoded by a stepwise UMI duplication via restriction and ligation. This enabled the analysis of mutational impact over the full GFP sequence, but multiple restriction and ligation steps render the protocol difficult and any imperfection (e.g. chimer formation in the ligation step, mutations in the restriction sites in the gene) scale badly with longer gene lengths, requiring more of those steps which might distort the sequence readout quantitatively and qualitatively.

A straightforward single read over the entire gene length would remove these ambiguities: the long primary nanopore sequencing readout does not need a processing assembly, but a simple comparison to their parent variant is sufficient. UMIC-seq is a more reliable and practically easier method in our opinion and results in a simple and scalable workflow, robust with regard to read coverage. By using true long reads, the workflow is the same for all gene lengths and the coverage over the length of the gene is uniform. Practically, the third-generation sequencing method we present here is easier to implement than any other: starting from a simple plasmid stock, the tagged sequencing library is obtained based on straightforward steps that only involve Gibson assembly and one PCR. Only a benchtop nanopore sequencer is required and the entire procedure can be carried out by one laboratory operator, who has direct control over the number of consensus reads (or UMIs) generated, as defined by the colony count. (All scripts and protocols can be found in the Supplementary Information and at https://github.com/fhlab/UMIC-seq).

We expect further applications of UMIC-seq beyond improvement of biocatalysts to arise e.g. in long-read deep mutational scanning, sequencing of immune repertoires and metagenomic samples or medical diagnostics. Applications involving larger sequence deviation between variants, such as metagenomics, might require the assembly of a reference sequence prior to polishing, which can be easily implemented in the pipeline with tools such as Miniasm or Racon. Additionally, deep mutational scanning (DMS) could seamlessly be integrated with UMIC-seq. DMS relates information on gene frequency changes to protein function for large scale studies of sequence-function relationships. Currently, DMS is based on short reads, enabling the analysis of comprehensive single-site saturation libraries or multi-point mutants in short domains (as discussed above). These large-scale multi-point mutagenesis datasets alone have recently been shown to enable the construction of small 3D structures. However, determining mutational effects covering full gene lengths has been identified as a crucial next step in technology development towards larger structures. With UMIC-seq accurate long reads become possible, so that multi-point mutations, even if far apart and randomly distributed across a whole gene, can be unearthed: studying functional epistasis in protein evolution in this way could simultaneously enable the mutagenesis-based determination of full protein structures. A synthesis of DMS and UMIC-seq would start with sequencing of the naïve library using the UMIC-seq workflow, resulting in accurate full-length sequences and their corresponding UMIs. In a second step, short-read sequencing of only the UMI-tag would suffice to determine the frequencies of variants after a functional assay, deriving fitness scores that quantify the (currently unpredictable) interactions of simultaneously occurring mutations throughout the full gene length of many variants.

Large numbers of long reads, obtained two orders of magnitude cheaper and with higher accuracy than possible at present, will powerfully inform emerging sequence-function relationships. This high information content will propel conventional phylogeny-based approaches, but also enable more advanced machine-learning techniques that predict cooperative fitness effects arising from intra-gene epistasis. The combination of easy-to-generate, large and low-cost datasets will make the exploration of sequence space for the analysis and engineering of functional proteins more accessible and likely more successful.
Methods
Reagents. All chemicals and oligonucleotides were purchased from Sigma-Aldrich unless otherwise noted. Enzymes were purchased from New England Biolabs. DH10B was provided by Johnson Matthey. Nanopore flow cells and kits were obtained from Oxford Nanopore Technologies. The plasmids pASK-IBA36 + and pRSF-Duet1 were purchased from IBA Lifesciences and Novagen, respectively.

Mutagenesis. The amine dehydrogenase (AmDH, Supplementary Sequence 1) was cloned into pASK-IBA36b + at NcoI and XhoI sites and expressed in fusion with a C-terminal Strep-tag. A random mutagenesis library was generated by error-prone PCR, as described in the Gene-Morph II random mutagenesis kit manual (Agilent). A target concentration of 100 ng template (parental AmDH plasmid or plasmid pool) was amplified with the Mutazyme II low-fidelity polymerase for 25 cycles. The resulting PCR product was ligated into pASK-IBA36b + at NcoI and XhoI sites and used to transform E. coli clone 10 G ELITE cells (Lucigen), yielding more than 10^2 transforms at an average mutation rate of 3.5 mutations per gene. Individual variants chosen for characterization were generated with primers containing the desired point mutation, derived from sequencing and network building, by IVA cloning52. Point mutations were incorporated in the forward primer (Supplementary Table 1) for whole-plasmid PCR, in between a homologous recombination region (±15 bp, Tm = 60 °C) and a template binding region (Tm = 60 °C). The reverse primer fit the homologous recombination region but was extended until a total Tm of 60 °C was reached. Plasmid containing the reference gene was used at 10 ng as template for 22 PCR cycles with the mutagenic primers, using Q5 polymerase according to manufacturer’s specification in a 25 µL reaction. Template DNA was removed by addition of 1 µL DpnI (FastDigest, Thermo Fisher) to the PCR reaction at 37 °C for 30 min. Finally, 1 µL of the unpurified digested PCR product was used to transform 25 µL chemically competent E. coli cells (NEB 5a high efficiency).

Directed evolution by absorbance-activated droplet sorting (AADS). E. coli cells were transformed with the AmDH library and grown in LB medium supplemented with 100 µg/mL ampicillin at 37 °C and 200 rpm. Protein expression was induced with 200 ng/mL anhydrotetracycline at an optical density at 600 nm (OD600) of 0.4–0.8 and cultivation continued overnight at 20 °C. The expression culture was washed with 10 mM glycine-KOH buffer (pH 9.0) and diluted in buffer containing 25% (v/v) Percoll into an OD600 of 0.01. Single cells were co-encapsulated in 280 µL droplets with substrate and lysis mix (10 mM WST-1 (NBS Biologicals), 10 mM (R)-1-methyl-3-phenylpropylamine, 2 mM NAD^+), 5 µg/ml 1-methoxy-5-methylphenazinum methyl sulfite (mPMS), 1 µM l-lysine (Merck) and 0.8x Cell/ytic B in 100-mM glycine-KOH pH 9.0) using a flow-focusing chip of 80 µm height and 50 µm width (Supplementary Fig. 4). Droplet generation flow rates were 8 µL/min for the aqueous solutions and 30 µL/min for the oil phase (HFE-7500 (3 M Novec) containing 1% (w/v) 008-FluoroSurfactant (Ran Biotechnologies)). Droplets were incubated for 2 h at 22 °C to allow the reaction to proceed. After 2 h of incubation, 250 µL of droplets were injected into the sorting channel (Supplementary Fig. 4) at 100–120 Hz (2-3 double emboss per min spaced by 30–40 µL/min-HFE-7500) and sorted according to their absorbance at 455 nm. For each sample, a user defined threshold was set to select the best 1000 variants out of approximately one million droplets at 25% occupancy. Plasmid DNA from sorted droplets was recovered52 and used for the next round of mutagenesis and sorting, and as input for the nanopore sequencing.

Sequencing library preparation. Sorted variants were the input for sequencing library preparation in form of pooled plasmid stocks. A detailed description of the sample preparation procedure is available (Supplementary Protocol 1)a. Briefly, PCR products were screened by using Q5 High-Fidelity 2X Master Mix according to the manufacturer’s instructions. The forward primer contained an experiment specific barcode for each selection round (sequences taken from the PCR Barcoding Expansion Pack 1-96 (EXP-PBC096) to allow de-multiplexing) and the reverse primer was designed of the UMI (Supplementary Fig. 1). A total of 50 bp of 10x SYPRO orange was used in a thermal shift assay. Small scale bio-transformations were set up as 500 µL reactions to determine the total turnover number, with a final concentration of 0.1 µM pure enzyme and 50 mM 4-phenyl-2-butanone in 1 M NH4Cl/NH4OH pH 9.6. The cofactor (2 mM NADH) was recycled by addition of 10 mg/mL glucose dehydrogenase (GDH-101) and 100 mM glucose. After 72 h at 30 °C and 800 rpm, reactions were extracted with 1 mL ethyl acetate and analyzed by UBHC. Measurements were performed on a 1290 Infinity II LC System (Agilent), fitted with a XBridge Phenyl column (4.6x30 mm, Waters) using a water-acetonitrile gradient with 0.1% (v/v) trifluoroacetic acid (10% acetonitrile for 30 sec, increase to 90% acetonitrile within 90 s, hold for 30 s) at 20 °C and a flow rate of 0.3 mL/min. 1-methyl-3-phenylpropylamine eluted at 1.43 min and 4-phenyl-2-butanone at 1.93 min and were detected at 210 nm. Conversions were determined from observed peak areas and used for the calculation of total turnover numbers.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability. Sequence data generated in this study are deposited in the European Nucleotide Archive under the accession code PRJEB35468. All finalized consensus sequences in FASTA format as well as a data file containing the corresponding mutations and counts in each round are available as Supplementary Data Files. Other data are available from the authors upon request. Source data are provided with this paper.

Code availability. All scripts used for the UMIC-seq workflow and an analysis of example data are available at https://github.com/flhalab/UMIC-seq54. Scripts to recreate the sequence similarity networks of Fig. 3 and the UMI simulations of Supplementary Fig. 1 are available at https://github.com/flhalab/UMIC-seq/tree/master/figures.
61. Mazurenko, S., Prokop, Z. & Damborsky, J. Machine learning in enzyme engineering. *ACS Catal.* **10**, 1210–1223 (2020).
62. Kintses, B. et al. Picoliter cell lysate assays in microfluidic droplet compartments for directed enzyme evolution. *Chem. Biol.* **19**, 1001–1009 (2012).
63. Zurek, P. J., Knyphausen, P., Neufeld, K., Pushpanath, A. & Hollfelder, F. UMI-linked nanopore consensus sequencing (UMIC-seq) of highly similar gene variants. *Protoc. Exch.* https://doi.org/10.21203/rs.3.pex-1177/v1 (2020).
64. Zurek, P. J., Knyphausen, P., Neufeld, K., Pushpanath, A. & Hollfelder, F. UMI-linked consensus sequencing enables phylogenetic analysis of directed evolution. *Github fhlab/UMIC-seq* https://doi.org/10.5281/zenodo.4055319 (2020).

**Acknowledgements**

The authors thank the members of the Hollfelder group for helpful comments on the manuscript and Iustina Slabu (Johnson Matthey) for the measurement of samples determining the total turnover numbers. P.J.Z. is supported by an EU H2020 studentship in the ITN ES-Cat [722610], K.N. held an EU Marie Curie H2020 fellowship in conjunction with Johnson Matthey, F.H. is an H2020 ERC Advanced Investigator [695669].

**Author contributions**

P.J.Z., P.K., and F.H. designed research. P.J.Z. and K.N. performed experiments. P.J.Z. and F.H. wrote paper with input from all authors.

**Competing interests**

The authors declare no competing interests.

**Additional information**

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41467-020-19687-9.

**Correspondence** and requests for materials should be addressed to F.H.

**Peer review information** *Nature Communications* thanks Colin Jackson, Michelle O’Malley and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[Open Access: This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.]

© The Author(s) 2020