Supplementary Information

Where DNA data storage loses its bits

Hassan Khodaiemehr, Malihe Aliasgari and Chen Feng

Contents

Supplementary Note 1: Theory and proofs **1**

Supplementary Note 2: Computational cost **12**

Supplementary Note 3: Software **13**

Supplementary Note 4: The decoding replay **13**

Supplementary Note 5: Reproducibility **14**

Supplementary Tables **15**

Supplementary Figures **29**

Supplementary Note 1: Theory and proofs

Channel model. A strand is $X = (X_1, \ldots, X_N) \in \Sigma^N$, $\Sigma = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. A read is produced as follows, independently of X and of the other reads except through X. There are $N+1$ insertion slots; slot $i \in \{0, \ldots, N-1\}$ precedes base $i+1$ and slot N follows base N. Slot i receives J_i inserted bases, where either $J_i \sim \mathrm{Bernoulli}(g_i)$ (single-insertion model) or $\Pr(J_i = n) = (1-g_i)g_i^n$ (geometric model); each inserted base has law u on Σ. Base i is deleted with probability d_i; otherwise it is emitted through the survivor matrix $M_i = (c_i\,\mathbf{1} + s_i K)/(1-d_i)$, where $c_i = 1 - d_i - s_i$, $\mathbf{1}$ is the identity and K is a substitution kernel with zero diagonal and unit row sums. Let A denote the alignment of a read: the insertion counts $(J_0, \ldots, J_N)$ and the deletion indicators $(D_1, \ldots, D_N)$. Given A, the read Y consists of $\sum_i J_i$ inserted bases, i.i.d. with law u, interleaved at known places with the outputs of the surviving bases, each through its own M_i. A is independent of X, the read length $L = |Y| = N - \sum_i D_i + \sum_i J_i$ is a function of A, and A has independent coordinates, so $\mathrm{H}(A) = \sum_i \mathrm{H}(J_i) + \sum_i h_2(d_i)$ and the law of L is a convolution. The k reads of a strand are conditionally independent given X, with alignments $A_1, \ldots, A_k$ i.i.d.

This model contains the symmetric channel of our earlier work ($g_i \equiv \pi_{\mathrm{i}}$, $g_N = 0$, $d_i \equiv p_{\mathrm{d}}$, $s_i \equiv p_{\mathrm{s}}$, uniform u and K), and every fitted model in this paper. It excludes context-dependent errors (for example homopolymer-length errors of nanopore sequencing); for those, the rates measured on reads (main text) apply, the model-based bounds do not.

Domain and conventions. Throughout, $0 \le g_i < 1$, $d_i, s_i \ge 0$, $d_i + s_i \le 1$ and $d_i < 1$ (a position deleted with probability one carries no information and is removed from the strand); $0 \log 0 = 0$, and $W\mathrm{H}(w/W) = 0$ when $W = 0$. All entropies are in bits and per strand unless a quantity is written per nucleotide, e.g. $\delta = \Delta/N$. *Reference channel.* For Propositions 1 and 2 we use the constant-rate

channel with $g_i \equiv \pi_{\mathrm{i}}$ and $J_i \sim \mathrm{Bernoulli}(\pi_{\mathrm{i}})$ for the N slots $i = 0, \ldots, N-1$ preceding the bases, no terminal slot, $d_i \equiv p_{\mathrm{d}}$, $s_i \equiv p_{\mathrm{s}}$, uniform u and K. Its reads have length at most $2N$, so it has $(4^{2N+1}-1)/3$ outputs. It is a separate channel from the positional model, fitted to each dataset; no statement proved or computed for it is transferred to the positional model.

Run-ambiguity term. Fix an input law p on Σ and let X be i.i.d. p. For a maximal run R of base b occupying positions $a, \ldots, a+r-1$ ($r \geq 2$), let $e_0(i) = \Pr(J_i = 0)$ and $e_1(i) = \Pr(J_i = 1)$, and define the weights

$$w_j = d_j \prod_{l \in R,\, l \neq j} c_l \prod_{i=a}^{a+r-2} e_0(i), \qquad v_j = e_1(j-1) u_b \prod_{l \in R} c_l \prod_{i \in S_R,\, i \neq j-1} e_0(i), \qquad j \in R,$$

where $S_R = \{a-1, \ldots, a+r-2\}$ are the r slots preceding the positions of R (its own slots) and the interior slots are $a, \ldots, a+r-2$. Here w_j is the probability, given X, that position j of R is deleted, the other positions of R are emitted unchanged and no interior slot receives an insertion; v_j is the probability that all positions of R are emitted unchanged and exactly one own slot, the one preceding j, receives exactly one inserted base, equal to b, while the other own slots receive none. Let $W = \sum_j w_j$, $V = \sum_j v_j$, and $\mathrm{H}(\bar{w})$ the entropy of w/W. Define

$$\rho(p) = \mathbb{E} \sum_R \big[W_R \, \mathrm{H}(\bar{w}_R) + V_R \, \mathrm{H}(\bar{v}_R) \big],$$

the expectation being over the runs of X; for i.i.d. input, the probability that R is a maximal run of b at positions $a, \ldots, a+r-1$ is $p_b^r (1-p_b)^{\mathbf{1}[a>1] + \mathbf{1}[a+r-1<N]}$.

Lemma 1. $\mathrm{H}(A \mid X, Y) \geq \rho(p)$.

Proof. Let E be the complete marked error history of a read: the alignment A (insertion counts and deletion indicators) together with the letters of all inserted bases and the output of every surviving base. Given X, the read Y is a deterministic function $Y = \phi(X, E)$, and A is a function of E. For a run R of X, let Γ_R be the set of histories in which exactly one position of R is deleted, every other position of R is emitted unchanged and no interior slot of R receives an insertion, and Γ'_R the set in which every position of R is emitted unchanged and exactly one own slot of R receives exactly one inserted base, equal to b, the other own slots receiving none. The sets are disjoint for one run, and different runs involve disjoint positions and slots. Define an equivalence relation on histories: $e \sim e'$ if they agree everywhere except that, on runs R with $e \in \Gamma_R$ (respectively Γ'_R), the deleted position (respectively the slot of the inserted b) may differ. Moving the deletion within R leaves $r-1$ unchanged copies of b in the same place, and moving the inserted b to another own slot leaves $r+1$ consecutive copies of b in the same place, so $\phi(X, \cdot)$ is constant on each class. Let $G = G(X, E)$ name the class, retaining every mark of E other than these locations. Given X and G, the history is determined by the chosen location on each such run, and A determines that location; conversely A and G determine E. Because the events at different positions and slots are independent given X, $\Pr(E = e \mid X)$ factorizes over positions and slots, and within a class it is proportional, on each run, to w_j (deletion at j) or v_j (insertion before j), the remaining factors being common to the class. Hence, given (X, G), the locations on different runs are independent with laws $\bar{w}_R$ and $\bar{v}_R$. Since Y is a function of (X, G), $\mathrm{H}(A \mid X, Y) \geq \mathrm{H}(A \mid X, Y, G) = \mathrm{H}(A \mid X, G) = \mathbb{E} \sum_R [\mathbf{1}_{\Gamma_R} \mathrm{H}(\bar{w}_R) + \mathbf{1}_{\Gamma'_R} \mathrm{H}(\bar{v}_R)]$, and $\Pr(\Gamma_R \mid X) = W_R$, $\Pr(\Gamma'_R \mid X) = V_R$. The alignment A is not enlarged by this argument, so $\mathrm{H}(A)$ elsewhere is unchanged. □

For constant rates the forgotten locations are uniform and $\mathrm{H}(\bar{w}_R) = \mathrm{H}(\bar{v}_R) = \log_2 r$, which recovers the term ρ_N of our earlier analysis of the symmetric channel; the device is the one used in the small-deletion analyses of[1,2].

Sandwich theorem. For a 4×4 stochastic matrix M, an input law p and $m \geq 0$, let $I_p(M, m) = \mathrm{I}(V; Z_1, \ldots, Z_m)$ where $V \sim p$ and the Z_t are i.i.d. outputs of M given V ($I_p(M, 0) = 0$), and let $c(M, m) = \max_p I_p(M, m)$. Let $C_N^{(k)} = \sup \mathrm{I}(X; Y_1, \ldots, Y_k)$ over all laws of X on Σ^N.

Theorem 1 (Sandwich)**.** *For every $k \geq 1$ and every input law p on Σ,*

$$\sum_{i=1}^{N} \mathbb{E}\, I_p(M_i, m_i) - k\,\Delta(p) \;\leq\; C_N^{(k)} \;\leq\; \sum_{i=1}^{N} \mathbb{E}\, c(M_i, m_i), \qquad m_i \sim \mathrm{Bin}(k, 1 - d_i),$$

where $\Delta(p) = \mathrm{H}(A) - \mathrm{H}(L) - \rho(p) \geq 0$. For $k = 1$ the lower bound reads

$$C_N \;\geq\; \mathrm{H}(L) - \mathrm{H}(A) + \sum_{i=1}^{N} (1 - d_i)\, \mathrm{I}(p; M_i) + \rho(p).$$

Proof. Upper bound. For any law of X, $\mathrm{I}(X; Y^k) \leq \mathrm{I}(X; Y^k, A^k) = \mathrm{I}(X; Y^k \mid A^k)$, because A^k is independent of X. Given A^k, the inserted bases are independent of X, and the remaining read symbols are m_i conditionally independent outputs of M_i for each base X_i, with m_i the number of reads in which base i survives. The channel from X to Y^k given A^k is therefore a product over positions of the channels $X_i \mapsto Z_i^{m_i}$, whence $\mathrm{I}(X; Y^k \mid A^k = a) \leq \sum_i c(M_i, m_i(a))$. Averaging over A^k, $m_i \sim \mathrm{Bin}(k, 1 - d_i)$.

Lower bound. Let X be i.i.d. p. Since $L^k = (L_1, \ldots, L_k)$ is a function of A^k, $\mathrm{H}(Y^k) = \mathrm{H}(L^k) + \mathrm{H}(Y^k \mid L^k) \geq \mathrm{H}(L^k) + \mathrm{H}(Y^k \mid A^k)$. Given A^k, the inserted bases contribute $\mathbb{E} \sum_{j,i} J_{j,i} \mathrm{H}(u)$ and, as the X_i are independent, the surviving copies contribute $\sum_i \mathbb{E}\, \mathrm{H}(Z_i^{m_i})$. Next, $\mathrm{H}(Y^k \mid X) = \mathrm{H}(Y^k, A^k \mid X) - \mathrm{H}(A^k \mid X, Y^k) = \mathrm{H}(A^k) + \mathrm{H}(Y^k \mid X, A^k) - \mathrm{H}(A^k \mid X, Y^k)$, and $\mathrm{H}(Y^k \mid X, A^k) = \mathbb{E} \sum_{j,i} J_{j,i} \mathrm{H}(u) + \sum_i \mathbb{E}\, \mathrm{H}(Z_i^{m_i} \mid X_i)$. Subtracting, $\mathrm{I}(X; Y^k) \geq \mathrm{H}(L^k) - \mathrm{H}(A^k) + \sum_i \mathbb{E}\, I_p(M_i, m_i) + \mathrm{H}(A^k \mid X, Y^k)$. The pairs (A_j, Y_j) are conditionally independent given X, so $\mathrm{H}(A^k \mid X, Y^k) = \sum_j \mathrm{H}(A_j \mid X, Y_j) \geq k\rho(p)$ by Lemma 1; also $\mathrm{H}(L^k) = k\mathrm{H}(L)$ and $\mathrm{H}(A^k) = k\mathrm{H}(A)$. Finally $\Delta(p) \geq 0$ because $\mathrm{H}(A) - \mathrm{H}(L) = \mathrm{H}(A \mid L) \geq \mathrm{H}(A \mid L, X, Y) = \mathrm{H}(A \mid X, Y) \geq \rho(p)$. □

Remark 1 (What Δ measures)**.** *For a fixed input law the exact gain from revealing the alignments is $G_k(p) = \mathrm{I}(X; Y^k, A^k) - \mathrm{I}(X; Y^k) = \mathrm{I}(X; A^k \mid Y^k)$. For one read, $\Delta(p) - G_1(p) = \mathrm{I}(A; Y \mid L) + \mathrm{H}(A \mid X, Y) - \rho(p) \geq 0$, so Δ is an upper bound on the synchronization loss of the chosen input, not its exact value; we call it the synchronization-loss bound. The gap between the two sides of Theorem 1 at $k = 1$ is $\Delta(p)$ plus the difference between $\sum_i (1 - d_i) I_p(M_i, 1)$ and $\sum_i \mathbb{E}\, c(M_i, m_i) = \sum_i (1 - d_i)\, c(M_i, 1)$, which vanishes when one p is optimal for every M_i (for example uniform p and doubly symmetric M_i). The upper bound uses $\sum_i \mathbb{E} \max_p$, which is valid and may be looser than $\sum_i \max_p \mathbb{E}$. The bound $k\Delta$ on the multi-read loss grows with k and is useful only for few reads, which is why multi-read rates are measured from data (main text). Both bounds are monotone in k after taking $\max_{j \leq k}$ on the left, since a receiver may ignore reads.*

Remark 2 (Computation)**.** *For Bernoulli insertions the read-length law has support of size $O(N)$ and $\mathrm{H}(L)$ is an exact convolution with $O(N^2)$ operations. For geometric insertions L has unbounded*

support; we compute the law $\tilde{L}$ of L conditioned on the event F that no slot has more than 12 insertions. With $g_{\max} = 0.123$ the largest fitted insertion parameter, $\Pr(F^c) \le \varepsilon := (N+1)g_{\max}^{13} \le 2.4 \times 10^{-10}$ for every fitted model ($N \le 158$). Since $\mathrm{H}(L \mid \mathbf{1}_F) \le \mathrm{H}(L) \le \mathrm{H}(L, \mathbf{1}_F)$,

$$|\mathrm{H}(L) - \mathrm{H}(\tilde{L})| \le h_2(\varepsilon) + \varepsilon\,[\mathrm{H}(\tilde{L}) + \mathrm{H}(L \mid F^c)].$$

Given F^c, $L \le N + \sum_s G_s$ has mean at most $N + 13 + (N+1)g_{\max}/(1-g_{\max}) \le 200$, so $\mathrm{H}(L \mid F^c) \le \log_2(201) + \log_2 e \le 9.1$ bits (maximum entropy of a non-negative integer with that mean), and $\mathrm{H}(\tilde{L}) \le 9.1$ likewise; hence $|\mathrm{H}(L) - \mathrm{H}(\tilde{L})| \le h_2(2.4 \times 10^{-10}) + 2.4 \times 10^{-10} \cdot 18.2 < 2 \times 10^{-8}$ bits per strand, far below the three-decimal precision at which the bounds are reported. Floating-point error in the convolution is below 10^{-12}. $\mathrm{H}(A)$ is in closed form ($h_2(g)/(1-g)$ per geometric slot). $\rho(p)$ is a sum of non-negative run-ambiguity terms truncated where the run probability falls below 10^{-14}; dropping non-negative terms can only decrease ρ and increase Δ, so the truncated lower bound remains a valid (slightly conservative) lower bound. $I_p(M, m)$ and $c(M, m)$ are evaluated on the $\binom{m+3}{3}$ composition classes of Z^m, which are sufficient statistics; c is obtained numerically by Blahut–Arimoto and reported through its dual bound $\max_x D(M_x^{(m)} \| q)$, an upper bound at every iteration (it is not a closed form), so every upper bound in which it enters, including the block cap of Proposition 1, is a certificate regardless of convergence; the primal value is used only to report the primal–dual gap.

Uniqueness and symmetry of the optimal input. Call the channel model above *symmetrized* if u is uniform and K is the uniform kernel $(J - \mathbf{1})/3$, with J the all-ones matrix; then $M_i = c_i'\mathbf{1} + s_i'(J - \mathbf{1})/3$ with $s_i' = s_i/(1 - d_i)$, and every base permutation applied position-wise preserves the channel. The fitted positional models are not symmetrized; Theorem 2 and the lengths $N^\star$ of Supplementary Table 4 refer to the symmetrized model with the same g_i, d_i, s_i, and permutation invariance is not claimed for the fitted models.

Lemma 2 (Uniform reads)**.** *For the symmetrized channel and X uniform on Σ^N, $\Pr(Y = y) = \Pr(L = |y|)\, 4^{-|y|}$.*

Proof. Given the alignment, each read symbol is either an inserted base (uniform, independent of everything else) or the output of a distinct uniform base X_i through the doubly stochastic M_i, which is again uniform; distinct symbols are functions of distinct independent uniform variables. (This concerns one read; reads of the same strand share their input and are not independent.) □

Theorem 2 (Uniqueness and symmetry)**.** *For the symmetrized channel let $P_N = \Pr(L = N)$,*

$$a = \Pr(\textit{no insertion and no deletion}) = \prod_i \Pr(J_i = 0) \prod_{i=1}^{N} (1 - d_i), \qquad \lambda = \prod_i \big(1 - \tfrac{4}{3} s_i'\big).$$

If $s_i' < \frac{3}{4}$ for all i and

$$a\lambda > P_N - a,$$

then for every $k \ge 1$ the map $P_X \mapsto \mathrm{I}(X; Y^k)$ is strictly concave, and so is $P_X \mapsto \sum_{k \ge 0} \pi(k) \mathrm{I}(X; Y^k)$ for every coverage law π with $\pi(\{1, 2, \ldots\}) > 0$. On every nonempty closed convex set $\mathcal{P}$ of laws on Σ^N that is invariant under base permutations (for example, homopolymer-constrained strands), each of these functions has a unique maximizer, and the maximizer is invariant under every permutation of the bases.

Proof. Let $Q(x,y) = W(y \mid x)$ for $|x| = |y| = N$. Split $Q = a\bigotimes_i M_i + R$: the first term collects the alignments with no insertion and no deletion, $R \geq 0$ the others with $|y| = N$. Every row of Q sums to P_N because L is independent of X; every column sums to P_N by the lemma, since $\sum_x Q(x,y) = 4^N \Pr(Y = y)$ for uniform X. As the M_i are doubly stochastic, rows and columns of R sum to $P_N - a$, and the Schur test gives $\|R\|_2 \leq P_N - a$. The eigenvalues of the symmetric M_i are 1 and $1 - \frac{4}{3}s_i' > 0$, so the smallest singular value of $a\bigotimes_i M_i$ is $a\lambda$, and by Weyl's inequality that of Q is at least $a\lambda - (P_N - a) > 0$. Hence $p \mapsto p^{\mathsf{T}}Q$ is injective, and so are the maps from the input law to the law of Y and to the law of Y^k, $k \geq 1$. With $\mathrm{I}(X;Y^k) = \mathrm{H}(Y^k) - \sum_x p(x)\mathrm{H}(Y^k \mid X = x)$, strict concavity of entropy and injectivity give strict concavity for each $k \geq 1$; the term $k = 0$ is zero, so the weighted sum is strictly concave when π puts positive mass on some $k \geq 1$. (If $\pi(0) = 1$ the objective is identically zero.) The functions are continuous on the compact set $\mathcal{P}$, so a maximizer exists, and strict concavity makes it unique; a base permutation maps $\mathcal{P}$ and the objective to themselves, hence maps the maximizer to a maximizer, which must be itself. □

For the constant-rate channel the condition reduces to $a^N(1 + \lambda^N) > P_N$ of our earlier analysis.

Remark 3 (Invariance is not uniformity)**.** *Base permutations act on Σ^N without being transitive: they preserve the pattern of equal and unequal neighbours, so an invariant law may give different total mass to repeated and non-repeated pairs. For $N = 2$, one read and a deletion-only channel with deletion probability d, let $r = \Pr(X_1 = X_2)$ with mass spread evenly within the two classes. The output is the full strand with probability $(1-d)^2$, one uniformly chosen base with probability $2d(1-d)$, and empty otherwise; the single base is uniform and carries $1 + r$ bits about X. Hence*

$$\mathrm{I}_r = (1-d)^2\big[h_2(r) + 2r + (1-r)\log_2 12\big] + 2d(1-d)(1+r), \qquad r^* = \big(1 + 3\cdot 2^{-2d/(1-d)}\big)^{-1} > \tfrac{1}{4}.$$

At $d = \frac{1}{2}$, the uniform input gives 1.625 bits per strand and $r^* = 4/7$ gives 1.7018387 bits, which our orbit solver reproduces to ten digits (and 3.4650 against 3.4676653 at $d = 0.1$). The theorem therefore establishes uniqueness and invariance, not optimality of the uniform input; the uniform input is used in Theorem 1 and Proposition 2 only as one admissible input, which yields valid lower bounds.

A converse that captures insertions and deletions.

Proposition 1 (Block cap; reference channel)**.** *For the reference channel, whose slots each precede a base and which has no terminal slot, the read of an $(n+m)$-base strand is the concatenation of the reads of its first n and last m bases, each produced by the reference channel of that length, independently given the strand. Revealing the cut between the two parts cannot decrease the information, so $C_{n+m}^{(k)} \leq C_n^{(k)} + C_m^{(k)}$ [3,4], and for $N = qn + r$, $0 \leq r < n$, $C_N^{(k)} \leq qC_n^{(k)} + C_r^{(k)}$ ($C_0 = 0$).*

The remainder term is retained: for $N = 108$ and 158 ($r = 3$) it raises the cap by 0.0017 and 0.0002 bits per nucleotide over $C_5/5$; all other lengths are multiples of five. With a terminal slot, as in the positional model, the factorization would need boundary-specific capacities, which is why the proposition is stated for the reference channel. Unlike the genie of Theorem 1, which reveals the whole alignment, this cap reveals only block boundaries, so the cost of locating insertions and deletions inside a block is retained. We evaluate C_n, $n \leq 5$, by Blahut–Arimoto restricted to input laws that are uniform within each orbit O of the base permutations, which loses nothing by symmetry of the channel and concavity. The reduced objective keeps the information within orbits:

$$F(z) = \mathrm{H}\big(\textstyle\sum_O z_O \bar{W}_O\big) - \sum_O z_O \bar{h}_O, \qquad \bar{W}_O = \frac{1}{|O|}\sum_{x\in O} W_x, \quad \bar{h}_O = \frac{1}{|O|}\sum_{x\in O}\mathrm{H}(W_x),$$

which we implement by computing one representative row per orbit, symmetrizing the output law over the 24 permutations, and updating orbit masses with $D(W_x\|q)$ for the representative x (equal within an orbit). The reported value is the dual bound $\max_x D(W_x\|q)$ over all original rows. Checks: the noiseless channel gives $C_2 = 4$ and $C_3 = 6$ bits; the two-base deletion channel of the remark above gives 1.7018387305 bits at $d = \frac{1}{2}$. At $n = 5$ there are 51 orbits and $(4^{11} - 1)/3 \approx 1.4 \times 10^6$ outputs.

Certified rates. For the reference channel (Bernoulli insertions in the N slots preceding the bases, no terminal slot, uniform inserted and substituted bases) with i.i.d. input p, let $Z = -\log_2 \Pr(X \mid Y^k)$, so that $\mathrm{I}(X; Y^k) = \mathrm{H}(X) - \mathbb{E}Z$. Write $p_\mathrm{c} = 1 - p_\mathrm{d} - p_\mathrm{s}$ and $b = \max(p_\mathrm{c}, p_\mathrm{s}/3)$, the largest probability with which a surviving base is emitted as any given symbol.

Lemma 3 (Range)**.** *$0 \le Z \le f(X) + \sum_{r=1}^{k} \big(\log_2 B_b(|Y_r|) - \log_2 P(E_r)\big)$, where $f = -\log_2 p(X)$, E_r is the error pattern of read r, with I, D, S insertions, deletions and substitutions,*

$$P(e) = (\pi_\mathrm{i}/4)^I (1 - \pi_\mathrm{i})^{N-I} p_\mathrm{d}^D (p_\mathrm{s}/3)^S p_\mathrm{c}^{N-D-S},$$

$$B_b(m) = \sum_{\substack{0 \le D \le N \\ 0 \le I = D+m-N \le N}} \binom{N}{D}\binom{N}{I} (\pi_\mathrm{i}/4)^I (1 - \pi_\mathrm{i})^{N-I} p_\mathrm{d}^D b^{N-D}.$$

Proof. $W(y_r \mid x) \ge P(E_r)$ because E_r maps x to y_r. For every x', $W(y \mid x')$ is a sum over alignments compatible with $|y|$; an alignment is fixed by its insertion and deletion sets, each inserted base equals its output symbol with probability $\pi_\mathrm{i}/4$ times the slot factor, and each surviving base is emitted as its aligned output with probability at most b. Hence $W(y \mid x') \le B_b(|y|)$ and $\Pr(x \mid y^k) \ge p(x) \prod_r P(E_r)/B_b(|y_r|)$. □

Without the condition $b \ge p_\mathrm{s}/3$ the envelope fails: for $N = 1$, no insertions or deletions, $p_\mathrm{c} = 0.1$, $p_\mathrm{s} = 0.9$ and uniform input, a substitution gives $Z = 1.737$ bits while the envelope with p_c in place of b gives 0.415. Every fitted reference channel has $p_\mathrm{c} > 0.75 \gg p_\mathrm{s}/3$, so $b = p_\mathrm{c}$ and the reported intervals are unchanged. The bound is a sum of $f(X)$ and i.i.d. pattern terms with explicit laws, so the truncation tail $t(\tau) = \mathbb{E}(g - \tau)^+$ of the envelope g is computable exactly.

Proposition 2 (Certified rates)**.** *Let $Z_1, \ldots, Z_n$, $n \ge 2$, be i.i.d. copies of Z obtained by simulating the reference channel, τ a truncation level fixed from the explicit laws of the envelope before sampling, $T_t = \min(Z_t, \tau)$, $\bar{T}$ their mean, $\hat{V} = \frac{1}{n-1}\sum_t (T_t - \bar{T})^2$ and $\varepsilon(\alpha) = \sqrt{2\hat{V}\ln(2/\alpha)/n} + 7\tau\ln(2/\alpha)/(3(n-1))$. With probability at least $1 - \delta$,*

$$\mathrm{H}(X) - \bar{T} - \varepsilon(\delta/2) - t(\tau) \;\le\; \mathrm{I}(X; Y^k) \;\le\; \mathrm{H}(X) - \bar{T} + \varepsilon(\delta/2).$$

Proof. $\mathbb{E}T \le \mathbb{E}Z \le \mathbb{E}T + t(\tau)$ by the lemma. The empirical Bernstein inequality[5], applied to $T_t/\tau \in [0, 1]$ and to $1 - T_t/\tau$, each with failure probability $\delta/2$, bounds $\mathbb{E}T$ from above and below; a union bound gives joint coverage $1 - \delta$. □

The probability is over the simulation only: the intervals certify the information of the fitted reference channel, not of the physical channel or of the unknown true parameters. Each interval of Supplementary Table 4 has $\delta = 10^{-3}$; the J intervals of the table hold simultaneously with probability at least $1 - J\delta$. Endpoints above two bits per nucleotide would be loose, not inconsistent.

Achievability of measured rates.

Proposition 3. *Let P be the true law of a strand on Σ^N, W the true channel to its reads, V any auxiliary channel and $q_V(y) = \sum_x P(x)V(y \mid x)$. Suppose $q_V(Y) > 0$ and $V(Y \mid X) > 0$ almost surely under $P \times W$ and that $\log V(Y \mid X)/q_V(Y)$ is integrable. Then*

$$R(P, W, V) = \mathbb{E} \log_2 \frac{V(Y \mid X)}{q_V(Y)} = \mathrm{I}_{PW}(X; Y) - \mathbb{E}_Y D\big(P_{X|Y}(\cdot \mid Y) \,\|\, r_V(\cdot \mid Y)\big) \le \mathrm{I}_{PW}(X; Y),$$

with $r_V(x \mid y) = P(x)V(y \mid x)/q_V(y)$. If strands are used memorylessly (independent strands, each with law P and channel W), $\max(R, 0)$ is achievable with codewords drawn i.i.d. from P and the decoding metric V, the generalized mutual information at $s = 1$[6].

Proof. $R = \mathrm{H}(X) + \mathbb{E} \log_2 r_V(X \mid Y)$, and $\mathrm{H}(X \mid Y) = -\mathbb{E} \log_2 P_{X|Y}(X \mid Y) \le -\mathbb{E} \log_2 r_V(X \mid Y)$ by Gibbs' inequality, the difference being the expected divergence. □

The law in q_V must be the law of the measured strands. If a uniform reference is used while the true input is another law, the inequality can fail by any amount: for a noiseless channel, a deterministic input and a noiseless auxiliary channel with uniform reference, $R = 2N$ whereas $\mathrm{I} = 0$. Reads within a strand may be correlated; only the input law, the support and the memoryless sequence of strands must be correctly specified, not V. In the main text the designs of thirteen datasets pass the uniformity checks of Supplementary Table 12; for the four datasets with designs without repeated bases the rate is evaluated with the empirical first-order law of the designs as P. If an input class is absent from the designs (for example, strands with repeated bases), no reweighting can recover the channel's behaviour on it.

Pool. Strand m of a pool of M strands is read K_m times, with K_m i.i.d. with law π, independent of the data; the K_m reads are conditionally independent outputs of the strand channel. In the ordered-cluster model the receiver observes $(K_m, Y_m^{K_m})_{m \le M}$.

Proposition 4. *The ordered-cluster capacity per nucleotide is $C_{\mathrm{ord}} = \frac{1}{N} \max_p \sum_k \pi(k) \mathrm{I}_p(X; Y^k)$, and every rate achievable when the receiver sees only the unordered multiset of reads is at most C_{ord}. If one input law P is used for all k, $L_0 = 0$ and $L_j \le \mathrm{I}_P(X; Y^j)$ for all j, then $\frac{1}{N} \sum_k \pi(k) \max_{j \le k} L_j \le C_{\mathrm{ord}} \le \frac{1}{N} \sum_k \pi(k) \overline{C}_k$ for upper bounds $\overline{C}_k \ge C_N^{(k)}$.*

The first statement is the clustering-genie argument of[7–9] with π in place of the Poisson law; the lower end follows from data processing, since a receiver may ignore reads, and needs no optimality of P. It is a lower bound on C_{ord} only; no lower bound on unordered capacity follows from it. If the L_j are confidence limits, the bound holds on the event that they hold simultaneously, which we ensure by a Bonferroni allocation over all measured $(j, \text{receiver})$ pairs; receivers (consensus or joint) are compared on the same strands, and the maximum is taken over limits, not over point estimates. The coverage law is unbounded; the measured rates for j beyond the largest measured k are taken equal to the last measured value (lower end) and the upper bounds to two bits (upper end). If a latent strand property Z affects both abundance and errors, the correct average is $\mathbb{E}_Z \sum_k \pi_c(k \mid Z) J_k(Z)$; we use $\sum_k \pi_c(k) \mathbb{E}_Z J_k(Z)$, justified by the weak association between read count and single-read score ($|\rho| \le 0.10$, Supplementary Table 12). In the model of[9] the unknown order costs a further $(1 - \pi(0)) \log_2 M/N$ bits per nucleotide; its converse is established there for channels without insertions and deletions and under asymptotic conditions, so we report index-adjusted values only as estimates.

Sharper bounds through blocks. The results of this paragraph hold for the channel model above with an arbitrary law of finite mean for each J_i (Bernoulli and geometric in this paper), and make no symmetry assumption unless one of the following hypotheses is named. *Symmetrized*: as above. *Stationary*: d_i, s_i and the law of J_{i-1} do not depend on $i \in \{1, \dots, N\}$ (the law of J_N may differ). *Constant-rate symmetric*: stationary and symmetrized with $J_i \sim \text{Bernoulli}(\pi_\text{i})$ for $i < N$ and $J_N = 0$, which is the reference channel.

Blocks. For consecutive blocks $I_1, \dots, I_m$ of positions, each block owns its bases and the slots that precede them, and the last block also owns slot N. Y_t is the output of block t, $L_{I_t} = |Y_t|$, $B = (L_{I_1}, \dots, L_{I_m})$, and S_c is the read position of the cut c between blocks t and $t+1$. Given X the pairs (A_{I_t}, Y_t) are independent, $Y = Y_1 \cdots Y_m$, (Y, B) determines every Y_t, and the L_{I_t} are independent of each other and of X. A *window* is a block of consecutive bases with the slots that precede them. For a block or window J with input law P_J, $h_P(J) = \mathrm{H}(A_J \mid X_J, Y_J)$ is its alignment equivocation; since $\mathrm{H}(A_J, Y_J \mid X_J) = \mathrm{H}(A_J) + \mathrm{H}(Y_J \mid X_J, A_J)$,

$$h_P(J) = \mathrm{H}(A_J) + \sum_{\text{slots } i \in J} \mathbb{E}[J_i]\,\mathrm{H}(u) + \sum_{\text{bases } i \in J} (1 - d_i)\,\mathbb{E}_{P_J}\mathrm{H}\big(M_i(X_i, \cdot)\big) - \mathbb{E}_{P_J}\mathrm{H}(Y_J \mid X_J = x), \quad (1)$$

which is linear in P_J and needs only the output laws of single inputs of length $|J|$.

Lemma 4 (Genie identity)**.** *For every law of X and $k \geq 1$ reads,* $\mathrm{I}(X; Y^k) = \mathrm{I}(X; Y^k \mid A^k) - \mathrm{I}(X; A^k \mid Y^k)$. *For every block partition,* $\mathrm{I}(X; Y) = \mathrm{I}(X; Y \mid B) - \mathrm{H}(B \mid Y) + \mathrm{H}(B \mid X, Y)$.

Proof. A^k and B are independent of X, so $\mathrm{I}(X; Y^k) = \mathrm{I}(X; Y^k, A^k) - \mathrm{I}(X; A^k \mid Y^k) = \mathrm{I}(X; Y^k \mid A^k) - \mathrm{I}(X; A^k \mid Y^k)$, and the same with B in place of A^k, using $\mathrm{I}(X; B \mid Y) = \mathrm{H}(B \mid Y) - \mathrm{H}(B \mid X, Y)$. □

Theorem 1 is this identity with the last term bounded: $0 \leq \mathrm{I}(X; A^k \mid Y^k) \leq k\Delta(p)$.

Remark 4 (Symmetrized channel)**.** *If the channel is symmetrized and X is uniform on Σ^N, then given (A, L) the read is uniform on Σ^L (each read symbol is an inserted base or the output of a uniform base through a doubly stochastic M_i), so Y is independent of A given L,* $\mathrm{H}(A \mid Y) = \mathrm{H}(A \mid L)$, *and at* $k = 1$

$$\mathrm{I}(X; Y) = \sum_{i=1}^{N} (1 - d_i)\mathrm{I}(U; M_i) + \mathrm{H}(L) - \mathrm{H}(A) + \mathrm{H}(A \mid X, Y).$$

The only slack in the lower bound of Theorem 1 is then $\mathrm{H}(A \mid X, Y) - \rho(U)$, *which the following results reduce. Without symmetrization only* $\mathrm{H}(A \mid Y) \leq \mathrm{H}(A \mid L)$ *holds, and the identity becomes the lower bound of Theorem 1.*

Lemma 5 (Boundary chain)**.** *Let X have any law. For each cut c between I_t and I_{t+1} choose windows $J_c^- \subseteq I_t$ ending at c and $J_c^+ \subseteq I_{t+1}$ starting after c, and let $J_c = J_c^- \cup J_c^+$. Every $h(\cdot)$ is computed under the marginal of the input law on the corresponding block or window; on J_c this is the product of the marginals on J_c^- and J_c^+ when the blocks are independent. Windows of different cuts may overlap. Then*

$$\mathrm{H}(A \mid X, Y) = \sum_{t=1}^{m} h(I_t) + \mathrm{H}(B \mid X, Y), \qquad \mathrm{H}(B \mid X, Y) \geq \sum_{c} \big[h(J_c) - h(J_c^-) - h(J_c^+)\big],$$

and every bracket is non-negative.

Proof. Call a slot or a base a unit. The read is the concatenation of the strings emitted by the units in order. Given $X = x$ the emissions are independent, the emission of base i depends only on x_i and on the rates of base i, and A is in bijection with the vector of emission lengths. Fix (x, y) with $\Pr(Y = y \mid X = x) > 0$ and boundaries between units with read positions $R_1, \dots, R_{r-1}$ ($R_0 = 0$, $R_r = |y|$), which split the units into segments $\sigma_1, \dots, \sigma_r$. The event $\{R = \varrho, Y = y\}$ is the event that segment j emits $y_{(\varrho_{j-1}, \varrho_j]}$ for every j, so

$$\Pr(R = \varrho \mid X = x, Y = y) \propto \prod_{j=1}^{r} \Pr\big(\sigma_j \text{ emits } y_{(\varrho_{j-1}, \varrho_j]} \;\big|\; x\big). \tag{2}$$

Hence, given (X, Y), R is a Markov chain, and given R_{j-1} and R_{j+1} the position R_j has the law of the internal cut of the channel of $\sigma_j\sigma_{j+1}$ with input x and output $y_{(R_{j-1}, R_{j+1}]}$. *Identity.* S is a function of A, so $\mathrm{H}(A \mid X, Y) = \mathrm{H}(S \mid X, Y) + \mathrm{H}(A \mid X, Y, S)$. Given (x, y) and $S = s$, the same product structure shows that the A_{I_t} are independent with laws $\Pr(A_{I_t} \mid X_{I_t} = x_{I_t}, Y_t = y_{(s_{t-1}, s_t]})$. Since (X_{I_t}, Y_t) has the law of the channel of block t under the marginal input law, $\mathrm{H}(A \mid X, Y, S) = \sum_t h(I_t)$. Given L, B and S determine each other, so $\mathrm{H}(S \mid X, Y) = \mathrm{H}(B \mid X, Y)$. *Inequality.* For each (x, y), the chain rule, then conditioning on more variables, then the Markov property give $\mathrm{H}(S \mid x, y) = \sum_c \mathrm{H}(S_c \mid S_{<c}, x, y) \geq \sum_c \mathrm{H}(S_c \mid (S_{c'})_{c' \neq c}, x, y) = \sum_c \mathrm{H}(S_c \mid S_{c-}, S_{c+}, x, y)$, with $S_{c\pm}$ the neighbouring cuts (or 0 and $|y|$). Let E_c^- and E_c^+ be the read positions of the outer ends of J_c^- and J_c^+. Because $J_c^- \subseteq I_t$ and $J_c^+ \subseteq I_{t+1}$, the boundaries are ordered $S_{c-} \leq E_c^- \leq S_c \leq E_c^+ \leq S_{c+}$. Conditioning on $E_c^\pm$, then the Markov property of this refined chain, give $\mathrm{H}(S_c \mid S_{c-}, S_{c+}, x, y) \geq \mathrm{H}(S_c \mid E_c^-, E_c^+, x, y)$. By (2) the last posterior is that of the internal cut of the window channel J_c with input x_{J_c} and output $y_{(E_c^-, E_c^+]}$, and the true $E_c^\pm$ delimit the output of J_c. Averaging, $\mathrm{H}(S_c \mid E_c^-, E_c^+, X, Y)$ is the internal-cut equivocation of J_c under its marginal input law; by the identity applied to J_c with blocks $J_c^\pm$ it equals $h(J_c) - h(J_c^-) - h(J_c^+) \geq 0$. Each window enters only the conditioning of its own term, so overlaps between windows of different cuts do not matter. No step uses equal rates, a particular slot-count law, or uniform u or K. □

Lemma 6 (Crossing runs)**.** *For a maximal run R meeting two or more blocks, let κ_R (κ'_R) be the block containing the deletion of Γ_R (the inserted base of Γ'_R), with the law induced by $\bar{w}_R$ ($\bar{v}_R$), and let $H_{\mathrm{cross}}(P) = \mathbb{E}\sum_R [W_R \mathrm{H}(\kappa_R) + V_R \mathrm{H}(\kappa'_R)]$, the sum running over such runs. For every input law P: (i) $\mathrm{H}(B \mid X, Y) \geq H_{\mathrm{cross}}(P)$; (ii) $\rho(P) \leq \sum_t \rho_{I_t}(P) + H_{\mathrm{cross}}(P) \leq \sum_t h(I_t) + H_{\mathrm{cross}}(P)$, where ρ_{I_t} is the run term of Lemma 1 for the channel of block t alone.*

Proof. (i) Define G as in the proof of Lemma 1, forgetting locations only on crossing runs. Then Y is a function of (X, G), and B is a function of G and of the blocks κ_R, κ'_R of the forgotten locations, which are independent given (X, G). A cut lies between two adjacent bases, so at most one maximal run crosses it. Given G, the read positions of the cuts crossed by R determine on which side of each of them the forgotten deletion (insertion) lies, hence its block. So B determines the forgotten blocks given G, and $\mathrm{H}(B \mid X, Y) \geq \mathrm{H}(B \mid X, G) = H_{\mathrm{cross}}(P)$. (ii) A run inside one block contributes equally to $\rho(P)$ and to $\sum_t \rho_{I_t}(P)$, since its slots belong to that block. For a crossing run R with parts R_k in blocks t_k, the chain rule for the location = (block, position) gives $\mathrm{H}(\bar{w}_R) = \mathrm{H}(\kappa_R) + \sum_k \Pr(\kappa_R = k) \mathrm{H}(\bar{w}_R \mid \kappa_R = k)$. For $j \in R_k$, w_j is the weight of j for the run R_k of block t_k times a factor not depending on j, so the conditional law is $\bar{w}_{R_k}$. Moreover $\Gamma_R \cap \{\kappa_R = k\} \subseteq \Gamma_{R_k}$, so $W_R \Pr(\kappa_R = k) \leq W_{R_k}$. The same holds for insertions, because the slots preceding the bases of R_k belong to block t_k. Summing gives the first inequality; the second is Lemma 1 for each block. □

Theorem 3 (Block lower bound)**.** *For every input law and windows J_c as in Lemma 5,*

$$C_N \;\geq\; \mathrm{I}(X;Y \mid B) - \Big[\sum_{t=1}^{m} \mathrm{H}(L_{I_t}) - \mathrm{H}(L)\Big] + \max\Big\{\sum_c \big[h(J_c) - h(J_c^-) - h(J_c^+)\big],\, H_{\mathrm{cross}}\Big\}.$$

If the blocks are independent, $X_{I_t} \sim P_t$, then $\mathrm{I}(X;Y \mid B) = \sum_t \mathrm{I}_{P_t}(X_{I_t};Y_t)$. If instead each block is conditionally independent of the earlier blocks given a function ∂_t of $X_{I_{t-1}}$ (for example a Markov input whose order does not exceed the block lengths), then $\mathrm{I}(X;Y \mid B) \geq \sum_t \mathrm{I}(X_{I_t};Y_t \mid \partial_t)$.

Proof. By Lemma 4, $\mathrm{I}(X;Y) = \mathrm{I}(X;Y \mid B) - \mathrm{H}(B \mid Y) + \mathrm{H}(B \mid X,Y)$. Here $\mathrm{H}(B \mid Y) \leq \mathrm{H}(B \mid L) = \sum_t \mathrm{H}(L_{I_t}) - \mathrm{H}(L)$, and Lemmas 5 and 6(i) bound $\mathrm{H}(B \mid X,Y)$. Given (X,B) the block reads are independent, and Y_t depends only on X_{I_t} and L_{I_t}, which is independent of X and of the other blocks; hence $\mathrm{H}(Y \mid X,B) = \sum_t \mathrm{H}(Y_t \mid X_{I_t}, L_{I_t})$. With independent blocks also $\mathrm{H}(Y \mid B) = \sum_t \mathrm{H}(Y_t \mid L_{I_t})$. Under the Markov condition, $\mathrm{H}(Y \mid B) = \sum_t \mathrm{H}(Y_t \mid Y_{<t}, B) \geq \sum_t \mathrm{H}(Y_t \mid X_{<t}, B) = \sum_t \mathrm{H}(Y_t \mid \partial_t, L_{I_t})$. Finally $\mathrm{I}(X_{I_t};Y_t \mid \cdot, L_{I_t}) = \mathrm{I}(X_{I_t};Y_t \mid \cdot)$, because L_{I_t} is a function of Y_t and independent of the inputs. □

Proposition 5 (Comparison with Theorem 1)**.** *For $r \leq \mathrm{H}(A \mid X,Y)$ let $T_1[r] = \mathrm{I}(X;Y \mid A) + \mathrm{H}(L) - \mathrm{H}(A) + r$. This is the lower bound of Theorem 1 at $k = 1$ with $\rho(p)$ replaced by r, and it bounds $\mathrm{I}(X;Y)$ from below for every input law. Let $\rho_I = \sum_t h(I_t) + \sum_c [h(J_c) - h(J_c^-) - h(J_c^+)]$ and $\rho^+ = \sum_t h(I_t) + \max\{\sum_c [h(J_c) - h(J_c^-) - h(J_c^+)], H_{\mathrm{cross}}\}$. For every input law: (i) $\rho^+ \geq \max\{\rho, \rho_I\}$, hence $T_1[\rho^+] \geq T_1[\max\{\rho, \rho_I\}] \geq T_1[\rho]$; the inequality $\rho_I \geq \rho$ can fail. (ii) If the blocks are independent, the bound of Theorem 3 equals $T_1[\rho^+] + \sum_t \mathrm{I}(A_{I_t};Y_t \mid L_{I_t})$; the version without H_{cross} equals $T_1[\rho_I] + \sum_t \mathrm{I}(A_{I_t};Y_t \mid L_{I_t})$. (iii) For $k \geq 1$ reads, Theorem 1 holds with $\rho(p)$ replaced by ρ^+ for X i.i.d. p.*

Proof. $T_1[r] \leq \mathrm{I}(X;Y)$ follows from Lemma 4 with $\mathrm{H}(A \mid Y) \leq \mathrm{H}(A \mid L) = \mathrm{H}(A) - \mathrm{H}(L)$ and $\mathrm{H}(A \mid X,Y) \geq r$. (i) $\rho^+ \leq \mathrm{H}(A \mid X,Y)$ by Lemmas 5 and 6(i); $\rho^+ \geq \rho_I$ trivially, and $\rho^+ \geq \rho$ by Lemma 6(ii). An example with $\rho_I < \rho$ is in the remark below. (ii) Given A the read splits into independent block reads, so $\mathrm{I}(X;Y \mid A) = \sum_t \mathrm{I}(X_{I_t};Y_t \mid A_{I_t})$. Lemma 4 for block t gives $\mathrm{I}_{P_t}(X_{I_t};Y_t) = \mathrm{I}(X_{I_t};Y_t \mid A_{I_t}) - \mathrm{H}(A_{I_t} \mid Y_t) + h(I_t)$, and $\mathrm{H}(A_{I_t} \mid Y_t) = \mathrm{H}(A_{I_t}) - \mathrm{H}(L_{I_t}) - \mathrm{I}(A_{I_t};Y_t \mid L_{I_t})$. Sum over t and use $\mathrm{H}(A) = \sum_t \mathrm{H}(A_{I_t})$. (iii) In the proof of Theorem 1, $\mathrm{H}(A_j \mid X, Y_j) \geq \rho^+$ for each read. □

Corollary 1 (Uniform input, symmetrized channel)**.** *If the channel is symmetrized and the block inputs are uniform, then $\mathrm{I}(A_{I_t};Y_t \mid L_{I_t}) = 0$ (Remark 4 for each block). The bound of Theorem 3 then equals $T_1[\rho^+]$, and its version without H_{cross} equals $T_1[\rho_I]$. In the stationary case with $N = 5m$, windows of five bases and J_c split $2+3$, $\rho_I = (m-1)(h_5 - h_2 - h_3) + \sum_t h(I_t)$, where $h(I_t) = h_5$ except for the last block, which also owns slot N.*

Remark 5 (What is guaranteed and what is computed)**.** Valid lower bounds. *Theorem 3 (independent or Markov blocks), $T_1[\rho_I]$ and $T_1[\rho^+]$ hold for the channel of this note and every input law covered there; no symmetry is needed.* No worse than Theorem 1 at the same input. *$T_1[\max\{\rho, \rho_I\}]$, $T_1[\rho^+]$ and Theorem 3 with independent blocks are at least $T_1[\rho]$ by Proposition 5; maximizing over block-product inputs therefore never does worse than maximizing Theorem 1 over i.i.d. inputs. $T_1[\rho_I]$ alone, and Theorem 3 without H_{cross}, carry no such guarantee: for the binary deletion channel with $d = 0.05$, input* 0000*, blocks* $2+2$ *and J_c split* $1+1$*, $\rho = 0.3430 > \rho_I = 0.2850$, $\rho^+ = 0.3615$ and $\mathrm{H}(A \mid X,Y) = 0.3789$.* Strictly sharper on the fitted reference channels. *For the uniform input on the fourteen constant-rate*

symmetric reference channels of Supplementary Table 4, computed exactly from (1) *with blocks of five bases (a final block of three where* $N = 5m+3$*) and windows split* $2+3$*,* $T_1[\rho_I]$ *exceeds* $T_1[\rho]$ *by* 4×10^{-5} *to* 0.112 *bits per nucleotide (column "blocks" of Supplementary Table 4); the chain term is* 0.005–3.07 *bits per strand and exceeds* H_{cross}*, so there* $\rho_I = \rho^+$ *and* $\rho_I \geq \rho$ *also follows from Proposition 5(i). In every dataset the block lower bound lies inside the simulation-based confidence interval for the rate of the uniform input (Proposition 2), which it now meets from below without sampling. Further, by direct evaluation: for ten of these channels Theorem 3 with the capacity-achieving laws of isolated blocks of* 5–8 *bases exceeds the uniform value, by* 1×10^{-4} *to* 5.8×10^{-3} *bits per nucleotide; and on a non-symmetrized channel (*$N = 4$*, position-dependent rates, up to two insertions per slot, a slot after the last base, non-uniform* u *and input),* $T_1[\rho] = 0.683 < T_1[\rho_I] = 1.863 <$ *Theorem 3* $= 1.874 < \mathrm{I}(X;Y) = 1.933$*, with* $\sum_t \mathrm{I}(A_{I_t}; Y_t \mid L_{I_t}) = 0.0109$ *as in Proposition 5(ii). Since* $\mathrm{H}(B \mid Y)$ *is bounded by* $\mathrm{H}(B \mid L)$*, inputs that mark block ends gain nothing in these bounds.*

For a block I, $C(I)$ is the capacity of the channel of block I alone: its bases, the slots it owns, and its output Y_I.

Proposition 6 (Block cap and its converse)**.** *For every partition into blocks,*

$$\sum_{t=1}^{m} C(I_t) - \Big[\sum_{t=1}^{m} \mathrm{H}(L_{I_t}) - \mathrm{H}(L)\Big] \;\leq\; C_N \;\leq\; \sum_{t=1}^{m} C(I_t).$$

If the channel is stationary, let C_n° *and* L_n° *be the capacity and output length of a window of* n *bases. Then* $C_{n+m}^\circ \leq C_n^\circ + C_m^\circ$*,* $C_{n+m}^\circ \geq C_n^\circ + C_m^\circ - \mathrm{H}(L_n^\circ) - \mathrm{H}(L_m^\circ) + \mathrm{H}(L_{n+m}^\circ)$*,* $C_\infty = \lim_N C_N^\circ/N$ *exists,* $NC_\infty \leq C_N^\circ \leq NC_\infty + \mathrm{H}(L_N^\circ)$*, and* $C_N^\circ - \mathrm{H}(J_N) \leq C_N \leq C_N^\circ$*.*

Proof. Upper bound: by Lemma 4, $\mathrm{I}(X;Y) \leq \mathrm{I}(X;Y \mid B)$. $B_{-t} = (L_{I_s})_{s\neq t}$ is independent of (X, Y_t) and the Y_t are conditionally independent given X, so $\mathrm{I}(X;Y \mid B) \leq \sum_t \mathrm{I}(X_{I_t}; Y_t \mid L_{I_t}) = \sum_t \mathrm{I}(X_{I_t}; Y_t) \leq \sum_t C(I_t)$. Lower bound: take independent capacity-achieving inputs on the blocks; by Lemma 4, $\mathrm{I}(X;Y) \geq \sum_t C(I_t) - \mathrm{H}(B \mid Y) \geq \sum_t C(I_t) - \mathrm{H}(B \mid L)$. Stationary case: interior blocks are windows, so the two inequalities give the sub- and superadditivity of C°; by Fekete's lemma $C_\infty = \inf_N C_N^\circ/N$, and $a_N = C_N^\circ - \mathrm{H}(L_N^\circ)$ is superadditive with $a_N/N \to C_\infty$, so $a_N \leq NC_\infty$. Finally, the read of the full channel is the window read Y° followed by the J_N bases of slot N, a degradation of Y°, so $C_N \leq C_N^\circ$; and (Y, J_N) determines Y° while J_N and the bases of slot N are independent of (X, Y°), so $\mathrm{I}(X;Y) \geq \mathrm{I}(X;Y,J_N) - \mathrm{H}(J_N) = \mathrm{I}(X;Y^\circ) - \mathrm{H}(J_N)$. □

Proposition 1 is the upper bound for the reference channel ($J_N = 0$, so $C_N = C_N^\circ$) with q blocks of n bases and one of r.

Proposition 7 (Length truncation)**.** *Fix a block* I *and* T*, let* $\varepsilon = \Pr(L_I > T)$ *and let* W'_x *be the law of* Y_I *given* $X_I = x$ *and* $L_I \leq T$*. For every law* q' *on* $\Sigma^{\leq T}$*,*

$$C(I) \;\leq\; (1-\varepsilon)\max_x D(W'_x \| q') + 2\,\mathbb{E}\big[L_I \mathbf{1}\{L_I > T\}\big],$$

and every input law P *satisfies* $\mathrm{I}_P(X_I;Y_I) \geq (1-\varepsilon)\mathrm{I}_P(X_I;Y')$ *with* $Y' \sim W'_{X_I}$*.*

Proof. Use the dual bound $C(I) \leq \max_x D(W_x\|q)$ with $q = (1-\varepsilon)q'$ on $\Sigma^{\leq T}$ and $q(y) = \Pr(L_I = \ell)4^{-\ell}$ for $|y| = \ell > T$, a probability law. L_I is independent of X_I, so $W_x = (1-\varepsilon)W'_x$ on $\Sigma^{\leq T}$, which contributes $(1-\varepsilon)D(W'_x\|q')$; the outputs of length $\ell > T$ contribute $\Pr(L_I = \ell)[2\ell - \mathrm{H}(Y_I \mid X_I = x, L_I = \ell)] \leq 2\ell \Pr(L_I = \ell)$. For the second claim, $\mathrm{I}(X_I;Y_I) = \sum_\ell \Pr(L_I = \ell)\mathrm{I}(X_I; Y_I \mid L_I = \ell)$, and dropping the terms $\ell > T$ leaves $(1-\varepsilon)\mathrm{I}(X_I;Y')$. □

Theorem 4 (Tiling bound)**.** *Let $U(I) \geq C(I)$ be certified for every block I of at most K consecutive bases, for example by Proposition 7 with q' the Blahut–Arimoto output law. Define $V(0) = 0$ and $V(n) = \min_{1 \leq j \leq \min(K,n)} \big[V(n-j) + U(\{n-j+1, \ldots, n\})\big]$, where the block ending at N owns slot N. Then $C_N \leq V(N)$. In the stationary case one may take $U(I) = U^{\circ}_{|I|} \geq C^{\circ}_{|I|}$ for every block, including the last (Proposition 6). If u and K are invariant under a group $\mathcal{G}$ of relabellings of Σ and q' is $\mathcal{G}$-invariant, $D(W'_x \| q')$ depends only on the $\mathcal{G}$-orbit of x; for a symmetrized channel $\mathcal{G} = S_4$ and the computation runs over restricted growth strings.*

Proof. By Proposition 6, $C_N \leq \sum_t C(I_t) \leq \sum_t U(I_t)$ for every partition into blocks of at most K bases, and $V(N)$ is the minimum over such partitions. For $\sigma \in \mathcal{G}$ applied to every base, $W'_{\sigma x}(\sigma y) = W'_x(y)$, so $D(W'_{\sigma x} \| q') = D(W'_x \| q')$ when $q' \circ \sigma = q'$. □

Corollary 2 (Gap of the tiling bound)**.** *Let $\mathcal{P} = \{I_1, \ldots, I_m\}$ be a partition into blocks of at most K consecutive bases attaining the tiling bound, $V(N) = \sum_t U(I_t)$. Then*

$$0 \;\leq\; V(N) - C_N \;\leq\; \sum_{t=1}^{m} \big[U(I_t) - C(I_t)\big] + \sum_{t=1}^{m} \mathrm{H}(L_{I_t}) - \mathrm{H}(L).$$

The right inequality holds for every partition into blocks of at most K bases, so $V(N) - C_N$ is at most the minimum of the right side over such partitions.

Proof. The left inequality is Theorem 4. For any admissible partition, $V(N) \leq \sum_t U(I_t)$, with equality for an attaining one, and Proposition 6 applied to the same partition gives $C_N \geq \sum_t C(I_t) - [\sum_t \mathrm{H}(L_{I_t}) - \mathrm{H}(L)]$. Subtracting gives the right inequality. Every slot and every base is owned by exactly one block, so $L = \sum_t L_{I_t}$, with slot N counted in L_{I_m}; the L_{I_t} are independent because they depend on disjoint sets of independent slot counts and deletion indicators; hence $\sum_t \mathrm{H}(L_{I_t}) - \mathrm{H}(L) = \mathrm{H}(B) - \mathrm{H}(L) = \mathrm{H}(B \mid L) \geq 0$, B determining L. Each $U(I_t) - C(I_t) \geq 0$ by the choice of U. □

The first sum is the slack of the certified block bounds: for $U(I)$ given by Proposition 7, $U(I) - C(I) \leq (1-\varepsilon)\delta_I + 2\mathbb{E}[L_I \mathbf{1}\{L_I > T\}]$, where $\delta_I = \max_x D(W'_x \| q') - C'(I)$ is the duality gap of the output law q' and $C'(I)$ the capacity of W', since $C(I) \geq (1-\varepsilon)C'(I)$ by the second claim of Proposition 7; rounding adds its margin. In the stationary case with $U(I_m) = U^{\circ}_{|I_m|}$, the last term also contains $C^{\circ}_{|I_m|} - C(I_m) \in [0, \mathrm{H}(J_N)]$, which is 0 when $J_N = 0$. The second sum is $\mathrm{H}(B \mid L)$, the uncertainty about where the block boundaries fall in the read once its length is known, which bounds what the boundaries revealed by the block cap are worth.

Scope. All results of this paragraph hold for the channel of this note except where a hypothesis is named. Numbers computed from them for the constant-rate symmetric reference channel are directly comparable with Supplementary Table 4, which uses the same channel; they do not apply to the positional fitted model. For that model the theorems and propositions of this paragraph apply unchanged, but every term has to be recomputed from the per-position fit (the block channels at their positions, the fitted insertion law with slot N, and the fitted u and K); the S_4 reduction applies only if u and K are uniform.

Supplementary Note 2: Computational cost

Insertions and deletions couple every read symbol to every strand position. Under conditional independence of reads given the strand, the channel likelihood $W(y_1, \ldots, y_k \mid x) = \prod_r W(y_r \mid x)$ of a given

strand needs only k one-dimensional forward recursions. The expensive quantity is the joint output law $q(y_1, \ldots, y_k) = \sum_x P(x) \prod_r W(y_r \mid x)$ under the input law, which normalizes the posterior and enters every rate: the reads share the unknown strand, so it requires a $(k+1)$-dimensional alignment lattice of $\prod_r (|Y_r|+1) \times (N+1)$ states, and the cost grows by a factor of about N with every additional read (Supplementary Table 15 and Supplementary Fig. 2a). Exact capacities are worse: a block of n bases has 4^n inputs and $(4^{2n+1}-1)/3$ outputs; reducing the inputs to orbits of the base permutations leaves 51 orbits at $n = 5$ against 1.4×10^6 outputs (23 s), and $n = 6$ would need a $187 \times 2.2 \times 10^7$ transition matrix (Supplementary Fig. 2b). The paper therefore uses closed-form bounds ($O(N^2)$ per dataset), exact capacities only for blocks of five bases, measured rates through auxiliary models (one-dimensional recursions for one read and consensus outputs, the exact joint lattice for $k \leq 3$ reads), simulation-based certification, and per-read soft detection in the receiver (0.17 ms per read). All timings are for one core of a cloud virtual machine.

Supplementary Note 3: Software

The package `dnacap` (Python with C kernels, MIT licence) implements every step of the paper:

- `dnacap fit`: expectation–maximization fit of the positional channel from designs and reads (FASTA and a design-index/read table, or clusters in the common cluster format);
- `dnacap bounds`: Theorem 1 bounds for k reads, the uniqueness length of Theorem 2, and the block cap of Proposition 1;
- `dnacap air`: single-read achievable rate measured on reads (Proposition 3);
- `dnacap pool` and `dnacap cost`: pool-level rates for any coverage and abundance distribution, and the cost-optimal coverage as a function of the read/write cost ratio;
- `dnacap decode`: storage of a random file in an IRA code over the designs (coset construction) and its decoding from the reads by the iterative soft receiver or consensus baselines;
- wrappers for the published baselines (BMA-lookahead, Iterative reconstruction, abPOA, Trellis BMA, DNAformer).

A test suite on simulated reads checks that the fit recovers the generating rates, that the bounds are ordered and monotone, that the measured rate lies below the upper bound, and that a simulated file is decoded. The scripts that produce every number, table and figure of the paper from the public data are included.

Supplementary Note 4: The decoding replay

The benchmark reuses strands that were synthesized without our code. A message M, uniform on $\{0,1\}^K$, is encoded into $c(M) \in \{0,1\}^{2SN_{\mathrm{d}}}$; the designs X of the chosen strands, with law P_X and independent of M, are fixed; and the decoder is given $S = X \oplus c(M)$ (on non-pilot positions) together with the reads of X. *Equivalence for uniform designs.* $\Pr(S = s \mid M = m) = P_X(s \oplus c(m))$. If P_X is uniform, S is uniform and independent of M, and (M, S, X) has the law of a prospective system in which the encoder draws a uniform scrambling sequence S independently of the message, shares it

with the decoder (common randomness, a legitimate coding resource that need not be charged to the payload), and synthesizes $X = c(M) \oplus S$. The reads of X are then exactly the reads the prospective system would obtain, and decoding success has the same law. This is the random-scrambling argument of ref. 10. Pilot positions carry design bases that are known to the receiver; for uniform designs these are uniform known symbols, as in a prospective pilot scheme with pseudorandom pilot values. *General designs.* Let $n = 2SN_{\mathrm{d}}$ be the number of non-pilot design bits and K the message length, with c injective and M uniform and independent of X. Given $M = m$, S is a bijective function of X, so $\mathrm{H}(S \mid M) = \mathrm{H}(X)$ and $\mathrm{I}(M;S) = \mathrm{H}(S) - \mathrm{H}(X)$. Because conditioning on either independent summand cannot increase entropy, $\mathrm{H}(S) \geq \mathrm{H}(S \mid X) = \mathrm{H}(c(M)) = K$, and $\mathrm{H}(S) \leq n$; hence

$$\max\{0,\ K - \mathrm{H}(X)\} \ \leq\ \mathrm{I}(M;S) \ \leq\ \min\{K,\ n - \mathrm{H}(X)\}.$$

The right-hand side is an upper bound on the side information that the offset reveals; an estimate of $\mathrm{H}(X)$ from above (for instance from first-order statistics) does not turn it into a lower bound, and the leakage can be zero even when $n - \mathrm{H}(X)$ is large (binary X uniform on $\{00, 11\}$ and C uniform on the same set give S independent of M). For the designs of Organick et al. and Meiser et al., which never repeat a base, $\mathrm{H}(X) \leq S \log_2(4 \cdot 3^{N-1})$ by the support size, so at the replayed density of 1.70 bits per nucleotide without pilots the leakage is at least $1.70 - 1.589 = 0.111$ bits per nucleotide; the replay is therefore not equivalent to a prospective system at that density, which in any case exceeds the ceiling of such designs. For scrambled designs the lower bound is vacuous and the upper bound is small only if $\mathrm{H}(X)$ is close to n, which first-order statistics cannot certify (a strand law with long-range dependence can have uniform pair statistics). Equivalence therefore rests on the assumption, intended by the pseudorandom scrambling with which these designs were produced, that whole designs are uniformly distributed; we state it as an assumption, not as something the statistics establish. With pilots, the same identities hold conditionally on the public pilot values. Constraints such as GC balance or homopolymer limits are not imposed on the codewords of the coset code, so the replay says nothing about constrained codes. *Statistics.* In the exploratory runs each strand set has its own designs, code, interleaver and message, and trial t uses set t mod 5 with an independent read subsample; trials on one set share its code, so they are neither independent nor identically distributed for a population-level file error rate, and their adaptive stopping does not keep the nominal coverage (Methods). These runs are therefore descriptive. The inferential statements rest on the confirmation run, in which every trial is a new file (new code, interleaver and message) on one of the five design sets, with a fixed number of trials per point and paired receivers; given the design sets its trials are independent, and the per-set counts (Supplementary Tables 20 and 19) show how much the design sets differ. Generalization to newly drawn designs is not covered by a confidence statement.

Supplementary Note 5: Reproducibility

The Supplementary Software contains: the `dnacap` package (channel fitting, Theorem 1 bounds, orbit-reduced Blahut–Arimoto with dual certificates, certified rates, measured rates, pool and cost planning, decoding); the decoding-benchmark harnesses `dna_stats_run_v13` (with and without collapsing identical reads, the latter as run `v13b`) and `dna_stats_run_v14clus` (clustering check), with the five strand sets of each of the nine decoding datasets (designs, reads after the length filter, fitted model, code seed), the run script, the analysis script and the 22,660 benchmark trial records (with and without

collapsing identical reads) the 5,020 clustering-check records and the 8,550 records of the earlier run; and the scripts that produce every table and figure of this version, with their input data. Every number of the revised analyses (design law, reference shift, pool-gap decomposition, abundance models, simultaneous confidence limits, published-system audit, orbit-solver checks) is produced by a script listed with its command and output file in the manifest `MANIFEST.md`, which also records software versions, random seeds, dataset accessions, read-filtering counts per dataset and the hardware used for the benchmark (25 worker processes, 19-hour budget, resumable).

Supplementary Tables

Supplementary Table 1: What each reported quantity is. Channel or data, input law, probability space of any confidence statement, and status: proved (a theorem about a stated model), computed (deterministic numerics for a stated model), estimated (a statistic of reads), demonstrated (the outcome of a receiver on reads).

Quantity	channel / data	input law	probability space	status
Theorem 1 bracket	fitted positional model	uniform i.i.d. (lower), any (upper)	none	proved, computed
Uniqueness length $N^\star$	symmetrized positional model	any	none	proved, computed
Block cap $(qC_5 + C_r)/N$	reference channel	any (capacity)	none (dual bound)	proved, computed
Certified interval	reference channel	uniform i.i.d.	channel simulation	proved, estimated
Measured rate, k reads	real reads	law of the designs	read sampling, fixed designs	estimated
Pool lower limit	real reads, coverage law	law of the designs	simultaneous over k and receivers	estimated, ordered clusters
Pool upper limit	fitted positional model, coverage law	any	none	computed, ordered clusters
Index-adjusted values	as above	as above	none	estimate (converse not proved for IDS)
Published-system efficiency	proxy fitted models	any	none	estimate, proxy- and model-conditional
File error rate	real reads, replay	designs (uniform for equivalence)	strand sets, codes, read sampling	demonstrated
Cost optimum	pool lower limit	law of the designs	none	model-conditional prediction
Strand-length optimum	fitted model, decay extrapolation	uniform i.i.d.	none	model-conditional prediction

Supplementary Table 2: Units and overheads. B information bits stored in M strands with payload length N (primers excluded) and primer length P; R_{c} code rate, N_{d} non-pilot positions per strand.

Quantity	definition and included overheads
Capacities, bounds, Δ, ρ in theorems	bits per strand; divided by N when reported per nucleotide ($\delta = \Delta/N$)
Measured rates, pool limits	bits per payload nucleotide; index and pilot positions are payload
Payload density	$R_{\mathrm{payload}} = B/(MN)$
Synthesized density	$R_{\mathrm{synth}} = B/(M(N+P)) = \frac{N}{N+P} R_{\mathrm{payload}}$; used in Fig. 5b and the cost model ($P = 40$)
Index cost (estimate)	$(1 - \pi_c(0)) \log_2 M/N$ bits per payload nucleotide, subtracted once in Table 2 and in cost and design curves
Decoding density	$2R_{\mathrm{c}}N_{\mathrm{d}}/N$ bits per payload nucleotide; pilots counted as overhead, primers excluded, no index (ordered clusters)
Published densities R	as reported per payload nucleotide (primers excluded, address/index included); Ding et al.: 248 nt excluding primers and a 2-nt indicator
Coverage c	mean sequenced reads per synthesized strand (all reads); usable coverage cf

Supplementary Table 3: Public datasets used. Reads: reads assigned to designs and used (all, full-payload); $\Delta\ell$: held-out log-likelihood gain of the positional over the symmetric model (nats/read); abundance: coefficient of variation of read counts per design and fraction of designs without reads in the full sample.

Dataset	Source	designs	reads (full)	$\Delta\ell$	CV / dropout
Song 0 d	figshare 10.6084/m9.figshare.17193170	210,000	56,475 (55,006)	0.07	0.47 / 0.0%
Organick id20	github.com/uwmisl/data-nbt17 (id20)	607,150	60,000 (58,525)	0.16	0.50 / 1.9%
Meiser unaged	figshare 10.6084/m9.figshare.21070684	7,373	187,496 (138,321)	0.48	0.49 / 0.7%
Meiser repaired	figshare 10.6084/m9.figshare.21070684	7,373	73,812 (25,175)	1.48	0.68 / 0.6%
Meiser aged	figshare 10.6084/m9.figshare.21070684	7,373	27,066 (5,461)	0.69	1.80 / 1.1%
Song 28 d at 70 C	figshare 10.6084/m9.figshare.17193170	210,000	7,311 (2,544)	0.40	0.85 / 0.1%
Song 56 d at 70 C	figshare 10.6084/m9.figshare.17192639	210,000	2,178 (649)	0.35	1.86 / 0.1%
Song 70 d at 70 C	figshare 10.6084/m9.figshare.17192639	210,000	1,561 (448)	-0.32	2.23 / 0.2%
DNAformer Illumina	Zenodo 10.5281/zenodo.13896773	109,873	3,181,202 (3,181,202)	0.05	0.31 / 0.0%
Lietard capped	ENA PRJEB43002	19,794	148,020 (99,060)	0.46	1.40 / 0.2%
Lietard normal	ENA PRJEB43002	19,794	110,060 (65,816)	0.51	1.56 / 0.1%
Lietard spaced	ENA PRJEB43002	19,794	53,646 (31,957)	0.57	2.74 / 2.2%
Antkowiak File 1	figshare 10.6084/m9.figshare.c.5128901	16,383	400,000 (185,418)	0.34	1.02 / 0.0%
Antkowiak File 2	figshare 10.6084/m9.figshare.c.5128901	49,149	6,649 (3,247)	0.50	5.09 / 0.0%
Microsoft CNR	github.com/microsoft/clustered-nanopore-reads-dataset	10,000	269,709 (269,709)	2.31	0.70 / 0.2%
Chandak exp. 0	github.com/shubhamchandak94/nanopore_dna_storage_data	1,466	340,585 (293,026)	4.20	0.58 / 0.0%
DNAformer nanopore	Zenodo 10.5281/zenodo.13896773	109,864	1,264,593 (1,264,593)	2.34	0.41 / 0.0%

Organick: 6,000 randomly chosen designs of 607,150 in file id20, with all their reads among the first 7.2 million; Song: 8,000 randomly chosen designs of 210,000; Chandak: experiment 0 (1,466 designs). The abundance is estimated from read counts, which include Poisson sampling noise; abundance model A keeps this noise, which lowers both pool-level limits, and model B removes it (Methods and Supplementary Table 13).

Supplementary Table 4: Theory for the fitted channels. Reference channel (constant rates, Bernoulli insertions before each base, no terminal insertion, uniform inserted and substituted bases), parameters $(\pi_{\mathrm{i}}, p_{\mathrm{d}}, p_{\mathrm{s}})$: Theorem 1 lower bound for the uniform input; blocks: block lower bound $T_1[\rho_I]$ for the uniform input (Theorem 3, Corollary 1; blocks of five bases, windows split $2+3$, computed exactly; script `block_lb.py`); two-sided confidence interval for the rate of the uniform input (Proposition 2, $n = 2 \times 10^4$ simulated strands, joint confidence $1 - 10^{-3}$ per interval); block cap $(qC_5 + C_r)/N$ (Proposition 1, dual bound); genie bound B_1; two reads where computed (n = 3,000). All values in bits per nucleotide and conditional on the fitted parameters. $N^\star$: largest prefix length for which the condition of Theorem 2 holds for the symmetrized positional model with the fitted g_i, d_i, s_i (equal to N: holds over the full payload); it concerns uniqueness and permutation invariance, not uniformity, of the optimal input.

Dataset	N	$(\pi_{\mathrm{i}}, p_{\mathrm{d}}, p_{\mathrm{s}})$	Thm. 1	blocks	certified, 1 read	$(qC_5 + C_r)/N$	B_1	certified, 2 reads	B_2	$N^\star$
Song 0 d	158	(0.0002, 0.0011, 0.0020)	1.965	1.965	[1.961, 1.968]	1.971	1.973	–	1.996	158
Organick id20	110	(0.0011, 0.0072, 0.0074)	1.858	1.859	[1.854, 1.866]	1.895	1.911	–	1.983	110
Meiser unaged	110	(0.0007, 0.0069, 0.0079)	1.860	1.861	[1.855, 1.866]	1.893	1.908	[1.956, 2.001]	1.982	110
DNAformer Illumina	140	(0.0001, 0.0002, 0.0009)	1.986	1.986	[1.982, 1.989]	1.987	1.988	–	1.998	140
Meiser aged	110	(0.0010, 0.0246, 0.0147)	1.678	1.683	[1.675, 1.689]	1.773	1.818	–	1.961	65
Meiser repaired	110	(0.0026, 0.0393, 0.0292)	1.469	1.485	[1.478, 1.493]	1.616	1.687	–	1.915	18
Lietard capped	50	(0.0094, 0.0989, 0.0101)	1.272	1.310	[1.300, 1.321]	1.523	1.706	–	1.942	22
Lietard normal	50	(0.0102, 0.1006, 0.0208)	1.179	1.231	[1.222, 1.243]	1.444	1.623	–	1.903	19
Lietard spaced	50	(0.0036, 0.1084, 0.0168)	1.217	1.253	[1.242, 1.263]	1.462	1.636	–	1.912	28
Antkowiak File 1	60	(0.0319, 0.0997, 0.0181)	1.075	1.164	[1.158, 1.179]	1.413	1.644	[1.566, 1.662]	1.913	13
Microsoft CNR	110	(0.0196, 0.0225, 0.0193)	1.535	1.561	[1.553, 1.568]	1.708	1.788	[1.833, 1.898]	1.950	30
Chandak exp. 0	108	(0.0555, 0.0491, 0.0476)	1.023	1.135	[1.130, 1.147]	1.377	1.554	–	1.853	13
DNAformer nanopore	140	(0.0110, 0.0144, 0.0174)	1.652	1.665	[1.659, 1.672]	1.770	1.818	–	1.958	54
Song 28 d at 70 C	158	(0.0011, 0.0131, 0.0131)	1.762	1.765	[1.760, 1.772]	1.827	1.853	–	1.969	69

Supplementary Table 5: Joint soft decoding versus consensus. Measured AIRs (bits/nt) for $k = 2, 3$ reads, and the pool-level lower limit at mean coverage c with consensus only and with joint decoding of up to three reads.

	$k=2$		$k=3$		pool, $c=3$		pool, $c=5$	
Dataset	cons.	joint	cons.	joint	cons.	joint	cons.	joint
Song 0 d	1.965	1.993	1.993	2.000	1.764	1.773	1.905	1.911
Organick id20	1.868	1.976	1.987	1.990	1.676	1.697	1.825	1.837
Meiser unaged	1.877	1.982	1.986	1.999	1.589	1.614	1.796	1.816
Meiser repaired	1.541	1.841	1.814	1.873	0.938	0.991	1.233	1.304
Meiser aged	1.738	1.933	1.738	1.982	0.682	0.707	0.946	0.996
Song 28 d at 70 C	1.779	1.808	1.779	–	1.079	1.096	1.359	1.380
DNAformer Illumina	1.989	1.998	2.000	2.000	1.849	1.851	1.961	1.962
Lietard capped	1.326	1.736	1.573	1.901	1.136	1.309	1.373	1.583
Lietard normal	1.253	1.677	1.516	1.880	1.021	1.191	1.264	1.486
Lietard spaced	1.278	1.709	1.515	1.873	0.952	1.105	1.171	1.371
Antkowiak File 1	1.171	1.596	1.468	1.804	0.882	1.021	1.136	1.336
Antkowiak File 2	0.363	0.539	0.363	0.702	0.290	0.416	0.337	0.550
Microsoft CNR	1.600	1.820	1.855	1.944	1.509	1.573	1.714	1.768
Chandak exp. 0	1.212	1.526	1.535	1.737	1.182	1.291	1.405	1.493
DNAformer nanopore	1.723	1.894	1.929	1.991	1.692	1.751	1.861	1.906

Supplementary Table 6: Shared errors between reads of the same strand. Observed probability that an error event of one read recurs in another read of the same design, control probability for a read of another design at the same position, and the normalized excess.

Dataset	designs	observed	control	excess
Song 0 d	3000	0.025	0.000	2.5%
Organick id20	3000	0.017	0.006	1.1%
Meiser unaged	3000	0.008	0.006	0.2%
Meiser repaired	3000	0.043	0.036	0.7%
Meiser aged	1315	0.047	0.023	2.5%
Song 28 d at 70 C	477	0.523	0.001	52.2%
Song 56 d at 70 C	111	0.896	0.042	89.1%
Song 70 d at 70 C	54	0.820	0.000	82.0%
DNAformer Illumina	3000	0.000	0.002	-0.2%
Lietard capped	3000	0.124	0.090	3.8%
Lietard normal	3000	0.112	0.090	2.5%
Lietard spaced	3000	0.135	0.106	3.2%
Antkowiak File 1	3000	0.098	0.074	2.6%
Antkowiak File 2	567	0.208	0.107	11.3%
Microsoft CNR	3000	0.058	0.016	4.3%
Chandak exp. 0	1466	0.063	0.032	3.2%
DNAformer nanopore	3000	0.037	0.011	2.6%

Supplementary Table 7: Context dependence of errors. Deletion and substitution rates per base for isolated bases (run 1), bases in runs of two and in runs of three or more; in runs: fraction of payload bases in runs of length two or more; homopolymer insertions: fraction of inserted bases equal to an adjacent design base; deletions in runs: observed fraction of deletions falling in runs of length two or more versus the fraction predicted by the fitted position-only model. Organick and Meiser designs contain no homopolymers.

			deletion rate			substitution rate		homopolymer	deletions in runs
Dataset	reads	in runs	run 1	run 2	run 3+	run 1	run 3+	insertions	obs. / pred.
Song 0 d	20,000	43%	0.0012	0.0013	0.0015	0.0018	0.0021	43%	46% / 43%
Organick id20	20,000	0%	0.0071	–	–	0.0068	–	42%	–
Meiser unaged	20,000	0%	0.0063	–	–	0.0073	–	61%	–
Meiser aged	5,461	0%	0.0195	–	–	0.0133	–	43%	–
Song 28 d at 70 C	2,544	44%	0.0126	0.0130	0.0120	0.0092	0.0152	17%	44% / 44%
DNAformer Illumina	20,000	38%	0.0003	0.0002	0.0002	0.0009	0.0008	58%	36% / 38%
Lietard capped	20,000	36%	0.0941	0.0952	0.0907	0.0100	0.0165	4%	36% / 36%
Lietard normal	20,000	36%	0.0968	0.0952	0.0898	0.0185	0.0302	5%	35% / 36%
Antkowiak File 1	20,000	43%	0.0886	0.0914	0.0846	0.0175	0.0357	3%	43% / 43%
Microsoft CNR	20,000	43%	0.0168	0.0210	0.0423	0.0166	0.0185	48%	56% / 43%
Chandak exp. 0	20,000	43%	0.0387	0.0438	0.0650	0.0390	0.0486	43%	50% / 44%
DNAformer nanopore	20,000	38%	0.0134	0.0159	0.0217	0.0126	0.0171	38%	44% / 38%

Supplementary Table 8: Where the bits go (values of main-text Fig. 2). Model-based accounting, in bits per payload nucleotide, at mean coverage c (sequenced reads per strand), ordered clusters, 10^6 strands, abundance model A. Unread (uniform pool): $2e^{-c}$; uneven (excess unread strands): $2(\pi_c^{\mathrm{all}}(0) - e^{-c})$; errors (genie error loss): $2(1 - \pi_c^{\mathrm{all}}(0)) - U(c)$; full-read gap: $U(c) - U(cf)$; residual (genie-to-AIR): $U(cf) - L(c)$; index: $(1-\pi_{cf}(0))\log_2(10^6)/N$, charged on strands with at least one usable read (the coverage law of the lower limit and of the net density), so it is smaller than $(1 - \pi_c^{\mathrm{all}}(0))\log_2(10^6)/N$ where few reads are usable; net: $L(c)$ minus index. The terms sum to two by construction and depend on the order of the counterfactuals. U, L: upper and lower pool limits (Methods).

c	dataset	N	unread	uneven	errors	full-read gap	residual	index	net
3	DNAformer Illumina	140	0.100	0.047	0.002	0.000	0.000	0.132	1.719
3	Song 0 d	158	0.100	0.111	0.005	0.010	0.001	0.112	1.661
3	DNAformer nanopore	140	0.100	0.076	0.036	0.000	0.038	0.130	1.621
3	Microsoft CNR	110	0.100	0.188	0.046	0.000	0.094	0.155	1.418
3	Chandak exp. 0	108	0.100	0.195	0.092	0.067	0.255	0.152	1.139
3	Song 28 d at 70 C	158	0.100	0.125	0.031	0.582	0.067	0.077	1.019
3	Lietard capped	50	0.100	0.141	0.065	0.206	0.179	0.314	0.995
3	Antkowiak File 1	60	0.100	0.087	0.083	0.450	0.259	0.234	0.787
10	DNAformer Illumina	140	0.000	0.003	0.000	0.000	0.000	0.142	1.855
10	Song 0 d	158	0.000	0.019	0.001	0.001	0.000	0.125	1.854
10	DNAformer nanopore	140	0.000	0.007	0.003	0.000	0.012	0.142	1.836
10	Microsoft CNR	110	0.000	0.034	0.009	0.000	0.052	0.178	1.727
10	Chandak exp. 0	108	0.000	0.052	0.027	0.021	0.214	0.178	1.507
10	Song 28 d at 70 C	158	0.000	0.014	0.005	0.185	0.149	0.115	1.533
10	Lietard capped	50	0.000	0.026	0.010	0.042	0.120	0.387	1.415
10	Antkowiak File 1	60	0.000	0.007	0.007	0.101	0.228	0.320	1.336

Supplementary Table 9: Capacity bracket under the fitted model alone. Both ends of Theorem 1 evaluated under the same fitted positional channel (decoding-experiment fit, strand set 0), the same uniform input and the same coverage law (abundance model A over the read counts of the 7,500 decoding designs): upper end $\sum_k \pi_c(k)\frac{1}{N}\sum_i \mathbb{E}\, c(M_i, m_i)$ (dual values, plus two bits times the mass beyond 40 reads); lower end $\sum_k \pi_c(k)\max_{j\le k}[\frac{1}{N}\sum_i \mathbb{E}\, I_u(M_i, m_i) - j\delta]^+$, δ from Table 1. Ordered clusters, index not included. The coverage law differs from that of Fig. 5a (a subset of designs), so the levels are not comparable with it; the width is the quantity of interest. Script `model_bracket.py`.

	Song 0 d ($\delta = 0.009$)			DNAformer Illumina ($\delta = 0.002$)		
c	lower	upper	width	lower	upper	width
2	1.554	1.567	0.013	1.669	1.672	0.003
3	1.725	1.741	0.016	1.846	1.851	0.004
5	1.855	1.874	0.019	1.957	1.962	0.005
10	1.923	1.945	0.022	1.991	1.997	0.006

Supplementary Table 10: Coverage-tail corrections. Amount added to the upper limit (two bits per nucleotide times the coverage mass beyond 60 reads, so that it remains an upper bound) and to the lower limit (rate at 60 reads applied to the tail, valid because a receiver may ignore reads), in bits per nucleotide, at mean coverage c (abundance model A). Below $c = 20$ all corrections are below 2×10^{-3}.

	$c = 20$		$c = 30$		$c = 50$	
Dataset	upper	lower	upper	lower	upper	lower
Song 0 d	0.0001	0.0001	0.0015	0.0013	0.0108	0.0100
Organick id20	0.0001	0.0001	0.0014	0.0013	0.0116	0.0108
Meiser unaged	0.0002	0.0000	0.0019	0.0003	0.0153	0.0050
DNAformer Illumina	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DNAformer nanopore	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Microsoft CNR	0.0002	0.0005	0.0006	0.0011	0.0011	0.0017
Chandak exp. 0	0.0002	0.0003	0.0004	0.0009	0.0009	0.0058
Antkowiak File 1	0.0001	0.0000	0.0007	0.0000	0.0033	0.0005
Lietard capped	0.0016	0.0002	0.0074	0.0016	0.0282	0.0106

Supplementary Table 11: Coverage ledger of the decoded datasets. Mapped: reads assigned to designs per design in the analysed sample (Supplementary Table 3); length-accepted: mean reads per design within the design-free length window used by every receiver in the decoding benchmark (the coverage axis of Table 3 and Fig. 6); f_{len}: their ratio; full-length: fraction of mapped reads passing the design-assisted full-length criteria (the usable fraction f of Table 1, used by the pool-level lower limits). A benchmark coverage c corresponds to about c/f_{len} mapped reads per design under independent filtering; relative read savings between receivers are unchanged by this rescaling.

Dataset	mapped	length-accepted	f_{len}	full-length f
Song 0 d	7.06	7.05	1.00	0.97
DNAformer Illumina	28.95	28.95	1.00	1.00
Organick id20	10.00	9.97	1.00	0.98
Meiser unaged	25.43	23.39	0.92	0.74
DNAformer nanopore	11.51	11.51	1.00	1.00
Microsoft CNR	26.97	26.97	1.00	1.00
Chandak exp. 0	232.3	231.6	1.00	0.86
Antkowiak File 1	24.42	13.77	0.56	0.46
Lietard capped	7.48	4.00	0.54	0.67

Supplementary Table 12: Law of the synthesized designs, and abundance versus quality. For the 7,500 designs of the decoding datasets: frequency of equal adjacent bases (uniform law: 0.25); empirical first-order entropy H_1 (bits/nt; the law without repeated bases has $\log_2(4 \cdot 3^{N-1})/N = 1.589$ at $N = 110$); first-order entropy deficit $2 - H_1$ (bits/nt; a diagnostic, not a bound on the information that the coset offset reveals, Supplementary Note 4); single-read rate on 1,000 reads (one per design, length within 10% of N, in-sample positional model, bits/nt) with q_V under the uniform law and under the empirical first-order law of the designs, and their difference; Spearman correlation between the read count of a design and its single-read score, and the mean score in the lowest and highest quartiles of read count. Standard errors of the rates: 0.002–0.014. The cross-fitted single-read rate of the main text minus the difference gives approximately 1.486 (Organick) and 1.488 (Meiser) bits/nt for the actual designs; this transfers an in-sample, length-filtered shift to a cross-fitted estimate and is an approximation. The Spearman correlation and quartile scores are diagnostics; they do not bound the covariance neglected when quality and coverage are averaged separately (Methods).

Dataset	N	equal adjacent	H_1	deficit $2 - H_1$	rate, uniform ref.	rate, design law	difference	ρ	score low / high quartile
Song 0 d	158	0.248	2.000	0.000	1.966	1.959	0.007	-0.09	1.961 / 1.954
Organick id20	110	0.000	1.570	0.430	1.866	1.481	0.385	+0.03	1.489 / 1.493
Meiser unaged	110	0.000	1.561	0.439	1.868	1.476	0.392	+0.06	1.465 / 1.477
DNAformer Illumina	140	0.209	1.989	0.011	1.986	1.975	0.011	-0.10	1.975 / 1.976
DNAformer nanopore	140	0.208	1.989	0.011	1.708	1.698	0.011	-0.01	1.708 / 1.696
Microsoft CNR	110	0.250	2.000	0.000	1.608	1.608	-0.000	+0.01	1.606 / 1.609
Chandak exp. 0	108	0.251	1.999	0.001	1.199	1.194	0.005	+0.08	1.162 / 1.223
Antkowiak File 1	60	0.250	1.988	0.012	1.168	1.157	0.010	+0.06	1.127 / 1.156
Lietard capped	50	0.205	1.992	0.008	1.405	1.396	0.009	+0.01	1.398 / 1.409

Supplementary Table 13: Pool-level limits for ordered clusters: decomposition and sensitivity. f: fraction of reads covering the payload. At mean coverage c (sequenced reads per strand): lower limit L (point estimates of measured rates, model A), upper limit U, model part $U(cf) - L(c)$ and truncation part $U(c) - U(cf)$ of the width (bits/nt). conf.: change of L at $c = 3$ when point estimates are replaced by simultaneous 95% lower confidence limits. B−A: change of L and U at $c = 3$ when the empirical read-count mixture (model A) is replaced by the deconvolved gamma abundance law (model B). ‡Designs without repeated bases: L uses the uniform reference law and is not a valid lower limit for these designs (Supplementary Table 12).

		$c = 2$				$c = 3$				$c = 5$				sensitivity, $c = 3$	
Dataset	f	L	U	model	trunc.	L	U	model	trunc.	L	U	model	trunc.	conf.	B−A (L / U)
Song 0 d	0.97	1.589	1.604	0.002	0.014	1.773	1.784	0.001	0.010	1.911	1.916	0.001	0.005	-0.002	+0.105 / +0.105
Organick id20‡	0.98	1.512	1.540	0.016	0.012	1.697	1.718	0.012	0.009	1.837	1.849	0.007	0.004	-0.006	+0.067 / +0.067
Meiser unaged‡	0.74	1.384	1.571	0.012	0.176	1.614	1.754	0.008	0.132	1.816	1.889	0.005	0.068	-0.004	+0.044 / +0.048
DNAformer Illumina	1.00	1.672	1.672	0.000	0.000	1.851	1.851	0.000	0.000	1.962	1.962	0.000	0.000	-0.001	+0.018 / +0.018
DNAformer nanopore	1.00	1.546	1.596	0.050	0.000	1.751	1.789	0.038	0.000	1.906	1.929	0.023	0.000	-0.012	+0.047 / +0.047
Microsoft CNR	1.00	1.365	1.466	0.100	0.000	1.573	1.667	0.094	0.000	1.768	1.845	0.077	0.000	-0.014	-0.008 / -0.011
Chandak exp. 0	0.86	1.089	1.418	0.246	0.083	1.291	1.613	0.255	0.067	1.493	1.788	0.251	0.043	-0.039	+0.022 / +0.037
Antkowiak File 1	0.46	0.779	1.513	–	–	1.021	1.730	0.259	0.450	1.336	1.902	0.262	0.304	-0.011	+0.024 / +0.053
Lietard capped	0.67	1.060	1.486	0.181	0.246	1.309	1.694	0.179	0.206	1.583	1.866	0.157	0.127	-0.009	+0.151 / +0.204

Supplementary Table 14: File error rates of the exploratory decoding runs. Adaptive stopping; descriptive (the inferential tests are in Supplementary Table 20). For each dataset, receiver and setting (identical reads collapsed within a cluster: on/off): coverage at which the file error rate is below 10% with 99% confidence (rule in Methods), and failures/trials at each coverage of the grid. The 22,660 trials of both runs are in the source data with their strand set. Collapsing identical reads removed the failure of our receiver on one Lietard strand set, whose 44 reads included 40 copies of one off-target molecule; it made vote-based reconstruction worse on Illumina data, where identical reads are mostly correct copies, and DNAformer failed at every coverage tested with it on the DNAformer Illumina pool.

Dataset	receiver	collapse	$c_{10\%}$	coverage: failures/trials
Song 0 d	this work	on	3.5	2.5: 20/20; 2.75: 20/20; 3: 20/20; 3.25: 9/20; 3.5: 0/60; 3.75: 0/60; 4: 0/60; 4.5: 0/60; 5: 0/60; 5.5: 0/60
Song 0 d	Trellis BMA	on	3.75	3: 20/20; 3.25: 17/20; 3.5: 29/170; 3.75: 0/60; 4: 0/60; 4.25: 0/60; 4.5: 0/60; 5: 0/60; 5.5: 0/60; 6: 0/60
Song 0 d	abPOA	on	6.0	3: 20/20; 3.25: 20/20; 3.5: 20/20; 3.75: 20/20; 4: 20/20; 4.25: 20/20; 4.5: 17/20; 5: 10/20; 5.5: 30/190; 6: 0/60
Song 0 d	Iterative reconstr.	on	5.5	3: 20/20; 3.25: 20/20; 3.5: 20/20; 3.75: 20/20; 4: 19/20; 4.25: 15/20; 4.5: 11/20; 5: 22/150; 5.5: 0/60; 6: 0/60
Song 0 d	Trellis BMA	off	3.75	3: 20/20; 3.25: 18/20; 3.5: 30/210; 3.75: 0/60; 4: 0/60; 4.25: 0/60; 4.5: 0/60; 5: 0/60
Song 0 d	abPOA	off	4.0	3: 20/20; 3.25: 20/20; 3.5: 16/20; 3.75: 9/20; 4: 3/110; 4.25: 0/60; 4.5: 0/60; 5: 0/60
Song 0 d	Iterative reconstr.	off	4.0	3: 20/20; 3.25: 20/20; 3.5: 19/20; 3.75: 7/20; 4: 1/80; 4.25: 0/60; 4.5: 0/60; 5: 0/60
DNAformer Illumina	this work	on	3.0	2: 20/20; 2.25: 20/20; 2.5: 20/20; 2.75: 10/20; 3: 2/90; 3.25: 0/60; 3.5: 0/60; 4: 0/60; 4.5: 0/60; 5: 0/60
DNAformer Illumina	Trellis BMA	on	3.0	2: 20/20; 2.25: 20/20; 2.5: 20/20; 2.75: 12/20; 3: 4/130; 3.25: 0/60; 3.5: 0/60; 4: 0/60; 4.5: 0/60; 5: 0/60
DNAformer Illumina	abPOA	on	3.5	2.5: 20/20; 2.75: 20/20; 3: 16/20; 3.25: 12/50; 3.5: 1/80; 3.75: 0/60; 4: 0/60; 4.5: 0/60; 5: 0/60; 5.5: 0/60
DNAformer Illumina	Iterative reconstr.	on	3.25	2.5: 20/20; 2.75: 18/20; 3: 11/40; 3.25: 3/110; 3.5: 0/60; 3.75: 0/60; 4: 0/60; 4.5: 0/60; 5: 0/60; 5.5: 0/60
DNAformer Illumina	DNAformer	on	—	3.5: 20/20; 3.75: 20/20; 4: 20/20; 4.25: 20/20; 4.5: 20/20; 4.75: 20/20; 5: 20/20; 5.5: 20/20; 6: 20/20; 6.5: 20/20
DNAformer Illumina	Trellis BMA	off	3.0	2: 20/20; 2.25: 20/20; 2.5: 20/20; 2.75: 12/20; 3: 3/110; 3.25: 0/60; 3.5: 0/60; 4: 0/60
DNAformer Illumina	abPOA	off	3.25	2.5: 20/20; 2.75: 15/20; 3: 28/300; 3.25: 0/60; 3.5: 0/60; 3.75: 0/60; 4: 0/60; 4.5: 0/60
DNAformer Illumina	Iterative reconstr.	off	3.25	2.5: 20/20; 2.75: 17/20; 3: 15/150; 3.25: 0/60; 3.5: 0/60; 3.75: 0/60; 4: 0/60; 4.5: 0/60
Organick id20	this work	on	3.0	2: 20/20; 2.25: 20/20; 2.5: 20/20; 2.75: 7/20; 3: 0/60; 3.25: 0/60; 3.5: 0/60; 4: 0/60; 4.5: 0/60; 5: 0/60
Organick id20	Trellis BMA	on	4.0	2.5: 20/20; 2.75: 20/20; 3: 20/20; 3.25: 19/20; 3.5: 7/20; 3.75: 30/300; 4: 1/80; 4.5: 0/60; 5: 0/60; 5.5: 0/60
Organick id20	abPOA	on	5.5	3: 20/20; 3.25: 20/20; 3.5: 20/20; 3.75: 20/20; 4: 20/20; 4.25: 20/20; 4.5: 20/20; 5: 12/20; 5.5: 5/140; 6: 0/60
Organick id20	Trellis BMA	off	3.5	2.5: 20/20; 2.75: 20/20; 3: 20/20; 3.25: 9/30; 3.5: 2/90; 3.75: 0/60; 4: 0/60; 4.5: 0/60
Organick id20	abPOA	off	4.0	3: 20/20; 3.25: 20/20; 3.5: 20/20; 3.75: 13/20; 4: 3/110; 4.25: 0/60; 4.5: 0/60; 5: 0/60
Meiser unaged	this work	on	3.5	2.5: 20/20; 2.75: 20/20; 3: 19/20; 3.25: 8/20; 3.5: 2/90; 3.75: 0/60; 4: 0/60; 4.5: 0/60; 5: 0/60; 5.5: 0/60
Meiser unaged	Trellis BMA	on	4.75	4: 19/20; 4.25: 14/20; 4.5: 16/70; 4.75: 6/160; 5: 0/60; 5.25: 0/60; 5.5: 0/60; 6: 0/60; 6.5: 0/60; 7: 0/60
Meiser unaged	abPOA	on	5.5	4: 20/20; 4.25: 20/20; 4.5: 20/20; 4.75: 16/20; 5: 9/30; 5.25: 32/220; 5.5: 1/80; 6: 0/60; 6.5: 0/60; 7: 0/60
Meiser unaged	Trellis BMA	off	4.75	4: 17/20; 4.25: 9/20; 4.5: 29/300; 4.75: 0/60; 5: 0/60; 5.25: 0/60; 5.5: 0/60; 6: 0/60
Meiser unaged	abPOA	off	5.0	4: 20/20; 4.25: 19/20; 4.5: 15/20; 4.75: 12/50; 5: 3/110; 5.25: 0/60; 5.5: 0/60; 6: 0/60

Dataset	receiver	collapse	$c_{10\%}$	coverage: failures/trials
DNAformer nanopore	this work	on	2.5	1.5: 20/20; 1.75: 20/20; 2: 20/20; 2.25: 11/20; 2.5: 0/60; 2.75: 0/60; 3: 0/60; 3.5: 0/60; 4: 0/60; 4.5: 0/60
DNAformer nanopore	Trellis BMA (pilots)	on	3.0	2: 20/20; 2.25: 20/20; 2.5: 18/20; 2.75: 29/300; 3: 0/60; 3.25: 0/60; 3.5: 0/60; 4: 0/60; 4.5: 0/60; 5: 0/60
DNAformer nanopore	abPOA (pilot-anchored)	on	4.25	3.5: 20/20; 3.75: 17/20; 4: 14/60; 4.25: 0/60; 4.5: 0/60; 4.75: 0/60; 5: 0/60; 5.5: 0/60; 6: 0/60; 6.5: 0/60
DNAformer nanopore	DNAformer	on	4.5	3.5: 20/20; 3.75: 20/20; 4: 13/20; 4.25: 17/150; 4.5: 0/60; 4.75: 0/60; 5: 0/60; 5.5: 0/60; 6: 0/60; 6.5: 0/60
DNAformer nanopore	Trellis BMA (pilots)	off	3.0	2: 20/20; 2.25: 20/20; 2.5: 17/20; 2.75: 25/300; 3: 0/60; 3.25: 0/60; 3.5: 0/60; 4: 0/60
Microsoft CNR	this work	on	3.5	2.5: 20/20; 2.75: 20/20; 3: 20/20; 3.25: 10/20; 3.5: 1/80; 3.75: 0/60; 4: 0/60; 4.5: 0/60; 5: 0/60; 5.5: 0/60
Microsoft CNR	Trellis BMA (pilots)	on	6.0	5: 17/20; 5.25: 11/20; 5.5: 15/60; 5.75: 23/300; 6: 0/60; 6.25: 0/60; 6.5: 0/60; 7: 0/60; 7.5: 0/60; 8: 0/60
Microsoft CNR	abPOA (pilot-anchored)	on	10.0	6: 20/20; 7: 20/20; 8: 19/20; 9: 21/300; 10: 0/60; 11: 0/60; 12: 0/60; 13: 0/60; 14: 0/60
Microsoft CNR	Trellis BMA (pilots)	off	6.0	5: 17/20; 5.25: 11/20; 5.5: 15/60; 5.75: 24/300; 6: 0/60; 6.25: 0/60; 6.5: 0/60; 7: 0/60
Chandak exp. 0	this work	on	2.75	2: 20/20; 2.25: 20/20; 2.5: 25/300; 2.75: 0/60; 3: 0/60; 3.25: 0/60; 3.5: 0/60; 4: 0/60; 4.5: 0/60; 5: 0/60
Chandak exp. 0	Trellis BMA (pilots)	on	3.25	2.5: 20/20; 2.75: 20/20; 3: 10/20; 3.25: 0/60; 3.5: 0/60; 3.75: 0/60; 4: 0/60; 4.5: 0/60; 5: 0/60; 5.5: 0/60
Chandak exp. 0	abPOA (pilot-anchored)	on	8.0	4.8: 20/20; 5.6: 20/20; 6.4: 20/20; 7.2: 7/20; 8: 0/60; 8.8: 0/60; 9.6: 0/60; 10.4: 0/60; 11.2: 0/60
Chandak exp. 0	Trellis BMA (pilots)	off	3.25	2.5: 20/20; 2.75: 20/20; 3: 10/20; 3.25: 0/60; 3.5: 0/60; 3.75: 0/60; 4: 0/60; 4.5: 0/60
Antkowiak File 1	this work	on	3.5	2.5: 20/20; 2.75: 20/20; 3: 20/20; 3.25: 15/20; 3.5: 0/60; 3.75: 0/60; 4: 0/60; 4.5: 0/60; 5: 0/60; 5.5: 0/60
Antkowiak File 1	Trellis BMA (pilots)	on	4.5	3.5: 20/20; 3.75: 20/20; 4: 12/20; 4.25: 19/300; 4.5: 0/60; 4.75: 0/60; 5: 0/60; 5.5: 0/60; 6: 0/60; 6.5: 0/60
Antkowiak File 1	abPOA (pilot-anchored)	on	9.0	5.4: 20/20; 6.3: 20/20; 7.2: 20/20; 8.1: 19/20; 9: 1/80; 9.9: 0/60; 10.8: 0/60; 11.7: 0/60; 12.6: 0/60
Antkowiak File 1	Iterative reconstr.	on	12.1	6.6: 20/20; 7.7: 20/20; 8.8: 20/20; 9.9: 17/20; 11: 10/150; 12.1: 0/60
Antkowiak File 1	Trellis BMA (pilots)	off	4.5	3.5: 20/20; 3.75: 20/20; 4: 13/20; 4.25: 23/300; 4.5: 0/60; 4.75: 0/60; 5: 0/60; 5.5: 0/60
Lietard capped	this work	on	2.75	2: 20/20; 2.25: 20/20; 2.5: 8/20; 2.75: 0/60; 3: 0/60; 3.25: 0/60; 3.5: 0/60
Lietard capped	Trellis BMA (pilots)	on	3.25	2: 20/20; 2.25: 20/20; 2.5: 20/20; 2.75: 19/20; 3: 28/300; 3.25: 0/60; 3.5: 0/60
Lietard capped	Trellis BMA (pilots)	off	3.25	2.5: 20/20; 2.75: 19/20; 3: 12/50; 3.25: 0/60; 3.5: 0/60; 3.75: 0/60

Supplementary Table 15: Cost of the computational kernels (one core of a cloud virtual machine). The joint output law of k reads under the uniform input requires a $(k+1)$-dimensional alignment lattice ($W(y \mid x)$: channel likelihood of one read for a given strand; columns 1–3 reads: output law q under the uniform input); exact capacity requires all 4^n inputs (orbits of base permutations) against $(4^{2n+1}-1)/3$ outputs.

a Output law of reads of a strand of length N

N	$W(y \mid x)$	1 read	2 reads	3 reads
50	0.010 ms	0.04 ms	0.002 s	0.14 s
100	0.025 ms	0.13 ms	0.027 s	8.02 s
150	0.099 ms	0.45 ms	0.225 s	–
200	0.210 ms	0.71 ms	0.812 s	–
300	0.417 ms	2.44 ms	2.244 s	–

b Exact block capacity C_n

n	orbits	outputs	time
1	1	21	0.0011 s
2	2	341	0.0017 s
3	5	5,461	0.0096 s
4	15	87,381	0.36 s
5	51	1,398,101	23 s

c Per read: EM E-step 0.38 ms ($N = 110$); soft detection 0.17 ms; joint detection of two reads 0.072 s.

Supplementary Table 16: Screening of receivers (earlier version; not the reported result). Historical 2-of-3 subsample screen: mean reads per strand at which a file stored in the same outer code over one set of 1,500 strands is recovered in at least two of three read subsamples (a descriptive screening statistic of an earlier version, not a reliability estimate and not the result reported in the main text; hard decisions calibrated with one symbol error rate). Columns under "2-of-3 screen (not the reported result)" are that historical screen for our iterative receiver and for published reconstruction methods (no pilots: our receiver without pilot bases, same density; =: array datasets, decoded without pilots throughout; DNAformer: better of soft output and calibrated hard decisions, Bar-Lev datasets only; Iterative reconstruction and BMA-lookahead: authors' implementations; abPOA: partial-order-alignment consensus; vote cons.: alignment-and-vote consensus); published methods: better of their pilot-free configuration and pilot-anchored reconstruction ($^{\text{P}}$: pilot-anchored better); — (x): not recovered up to x reads per strand (the largest coverage available or tested). The column Table 3 copies only the reported coverage of our receiver from the main-text benchmark (five strand sets, file error rate below 10% with 99% confidence) and supersedes the 2-of-3 "this work" cell. Density: information bits per nucleotide, identical for all receivers. Limit: smallest mean coverage at which the pool-level upper limit, and the measured-rate lower limit, reach the density.

		2-of-3 screen (not the reported result)								
Dataset	density	this work	no pilots	DNAformer	Iterative	abPOA	BMA-LA	vote cons.	Table 3	limit (UB / LB)
Song 0 d	1.80	3.5	=		4	4	5	4.5	3.5	3.2 / 3.3
DNAformer Illumina	1.80	3	=	4	3	3	3.5	3.5	3	2.7 / 2.7
Organick id20	1.70	3	=		4.5	4	6	4.5	3	2.9 / 3.1
Meiser unaged	1.70	3.5	=		6	5	10	7	3.5	2.7 / 3.7
DNAformer nanopore	1.44	2.5	3.5	4.5	6^{P}	4^{P}	7^{P}	8	2.5	1.6 / 1.7
Microsoft CNR	1.44	3.5	20		12	10^{P}	20	— (25)	3.5	2.0 / 2.3
Chandak exp. 0	0.96	2.5	— (30)		12	8^{P}	25^{P}	— (50)	2.75	1.0 / 1.6
Antkowiak File 1	1.12	3.5	— (12)		10	10^{P}	— (12)	— (12)	3.5	1.1 / 3.6
Lietard capped	1.12	3	3.5		— (3.5)	— (3.5)	— (3.5)	— (3.5)	2.75	1.2 / 2.3

Trellis BMA exponents (backward-pass, own-belief, other-reads and output exponents; Methods): $(1, 0, 0.25, 1)$ for Song, Organick, DNAformer Illumina and nanopore, Chandak and Lietard; $(1, 0, 0.25, 0.7)$ for Meiser and CNR; $(1, 0, 0.5, 0.7)$ for Antkowiak. The receivers carried into the benchmark of Table 3 were, for each dataset, ours, Trellis BMA and the best one or two published methods of this screening. Reported coverages are those of Table 3, not the 2-of-3 screen.

Supplementary Table 17: Ablation of the receiver, and anchoring without pilots. a, Historical 2-of-3 subsample screen of receivers built on the same detector and code (earlier version: mean reads per strand, on a half-read grid, at which the file is recovered in at least two of three read subsamples of one strand set; a descriptive comparison of the variants, not a reliability estimate and not the result reported in the main text): alignment-and-vote consensus with hard decisions, the same consensus passed through the soft detector, per-read soft detection without iterations with the code, and the iterative receiver. Columns under "2-of-3 screen (not the reported result)" are that historical screen; the last column copies the reported coverage of the iterative receiver from Table 3 (five strand sets, file error rate below 10% with 99% confidence) and supersedes the screening value, including the Chandak 2.5 and Lietard 3 cells of the 2-of-3 screen (reported: 2.75 and 2.75); ideal: reads grouped by their design; real: the binning released by the authors of the DNAformer data (design-free clustering of the other pools under the full statistical protocol: Supplementary Table 18). — (x): not recovered up to x. Densities include pilots. **b**, Symbol error rate (%) of the per-strand estimates before the outer code (product of per-read likelihoods, 300 strands) at mean coverage c, without pilots (uniform prior; leave-one-out iteration over the reads of a strand; prior centred on the consensus; joint detection of read pairs) and with pilots every tenth or fifth base (non-pilot positions).

			2-of-3 screen (not the reported result)				
Dataset	density	cluster	hard cons.	soft cons.	per-read soft	iterative	Table 3
Antkowiak File 1	1.12	ideal	— (12)	10	5	3.5	3.5
Chandak exp. 0	0.96	ideal	— (50)	5	3.5	2.5	2.75
Microsoft CNR	1.44	ideal	— (25)	6	4.5	3.5	3.5
DNAformer Illumina	1.80	real	3.5	3	3	3	3
DNAformer nanopore	1.44	real	8	3.5	3	2.5	2.5
Lietard capped	1.12	ideal	— (3.5)	— (3.5)	3.5	3	2.75
Meiser unaged	1.70	ideal	7	7	4.5	3.5	3.5
Organick id20	1.70	ideal	4.5	4.5	3.5	3	3
Song 0 d	1.80	ideal	4.5	4	3.5	3.5	3.5

b

			without pilots			with pilots	
Dataset	c	uniform	leave-one-out	consensus	pairs	every 10	every 5
Microsoft CNR	4	31.6	23.9	22.3	22.9	4.4	2.2
Microsoft CNR	6	30.9	20.6	20.0	20.5	3.2	1.5
DNAformer nanopore	4	19.5	10.0	8.9	11.1	1.8	1.1
DNAformer nanopore	6	17.9	6.5	4.9	6.6	1.1	0.5
Antkowiak File 1	4	40.7	34.9	38.9	35.7	23.1	11.3
Antkowiak File 1	6	38.8	29.8	36.5	32.4	20.7	7.9

Supplementary Table 18: Decoding with design-free clustering. **a**, decoding (exploratory protocol); **b**, pool-size stress test of the clustering. In every trial the reads of all 1,500 designs of the strand set were pooled, shuffled and clustered without labels (Methods); each design received the largest cluster whose majority read is its own, so lost and foreign reads are included but index recovery is not. Purity: fraction of reads in a cluster of their own design's majority; recall: fraction of reads that reach their design (means over all trials). Coverage (reads per strand) at which the file error rate is below 10% with 99% confidence, with failures/trials, with ideal clusters (Table 3) and with clustering, for our receiver and Trellis BMA, the best published method on these pools, each at the read-collapsing setting of Table 3. Same strand sets and statistical rule as the main benchmark (trial cap 150); 5,020 trials. The two DNAformer pools come clustered by their authors and are already in Table 3. ‡Designs without repeated bases.

			this work		Trellis BMA		saving	
Dataset	purity	recall	ideal	clustered	ideal	clustered	ideal	clustered
Song 0 d	1.000	1.000	3.5	3.5 (0/60)	3.75	3.75 (0/60)	7%	7%
Organick id20‡	0.998	0.981	3	3 (2/90)	3.5	3.75 (0/60)	14%	20%
Meiser unaged‡	0.999	0.984	3.5	3.75 (0/60)	4.75	5 (2/90)	26%	25%
Microsoft CNR	1.000	0.998	3.5	3.5 (4/130)	6	6 (2/90)	42%	42%
Chandak exp. 0	0.999	0.961	2.75	2.75 (2/90)	3.25	3.5 (1/80)	15%	21%
Antkowiak File 1	0.991	0.914	3.5	4.25 (0/60)	4.5	5.5 (5/140)	22%	23%
Lietard capped	0.999	0.962	2.75	3 (0/60)	3.25	3.5 (0/60)	15%	14%

b Pool-size stress test: the same clustering on one strand set and on the union of the five sets (distinct designs only), at the coverage of the ideal-cluster threshold, mean of two subsamples.

		one set (1,500)		five sets	
Dataset	designs (five sets)	purity	recall	purity	recall
Song 0 d	5,122	1.000	1.000	1.000	1.000
Organick id20	4,590	0.996	0.978	0.997	0.981
Meiser unaged	5,009	1.000	0.984	0.999	0.983
Microsoft CNR	5,590	1.000	0.998	0.999	0.985
Chandak exp. 0	1,466	0.999	0.953	0.999	0.953
Antkowiak File 1	6,243	0.994	0.906	0.988	0.852
Lietard capped	6,428	1.000	0.963	0.999	0.941

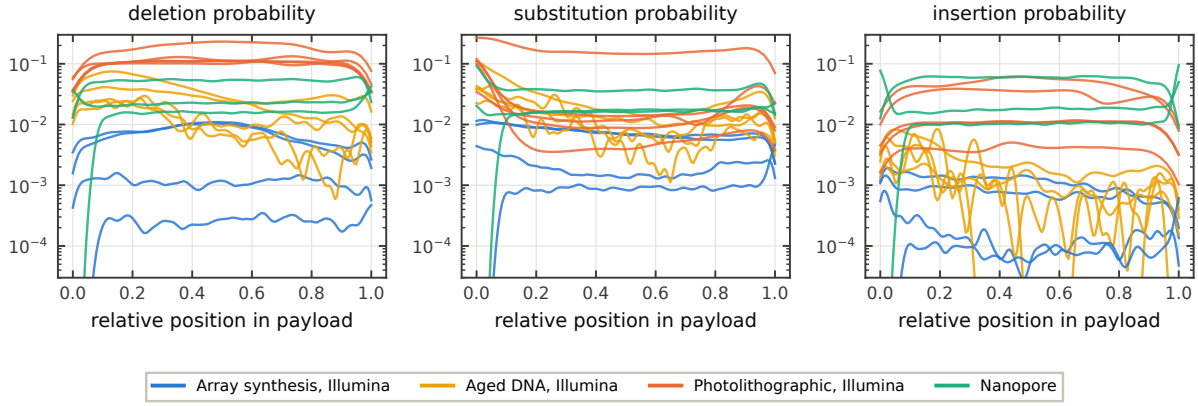
Supplementary Table 19: Failures per strand set at the first certified passing points (exploratory runs). For each dataset of Table 3: our receiver at its first certified passing point and one grid step below, and the best published method at its first certified passing point ((d): with identical reads collapsed); failures/trials in total, the exact 99% upper confidence limit of the file error rate treating trials as independent, failures/trials on each of the five strand sets, and the largest per-set failure fraction. Trials on the same set share its designs and code and were stopped adaptively, so these intervals are descriptive (Supplementary Note 4; the inferential tests are in Supplementary Table 20); the per-set counts show where failures concentrate. As the stopping rule implies, several reported thresholds of both receivers lie close to the target: upper limits of 0.089–0.099 for our receiver on DNAformer Illumina, Meiser and Microsoft CNR, and 0.094–0.099 for Trellis BMA on DNAformer Illumina, Organick and Meiser; at the same coverages most strand sets have no failure.

Dataset	receiver	c	total	99% UL	set 0	set 1	set 2	set 3	set 4	worst set
Song 0 d	this work	3.5	0/60	0.085	0/12	0/12	0/12	0/12	0/12	0%
	this work (below)	3.25	9/20	0.743	0/4	2/4	2/4	2/4	3/4	75%
	Trellis BMA (d)	3.75	0/60	0.085	0/12	0/12	0/12	0/12	0/12	0%
DNAformer Illumina	this work	3	2/90	0.099	0/18	0/18	0/18	2/18	0/18	11%
	this work (below)	2.75	10/20	0.782	2/4	1/4	1/4	3/4	3/4	75%
	Trellis BMA (d)	3	4/130	0.094	0/26	1/26	0/26	3/26	0/26	12%
Organick id20	this work	3	0/60	0.085	0/12	0/12	0/12	0/12	0/12	0%
	this work (below)	2.75	7/20	0.657	1/4	3/4	0/4	2/4	1/4	75%
	Trellis BMA	3.5	2/90	0.099	0/18	1/18	0/18	0/18	1/18	6%
Meiser unaged	this work	3.5	2/90	0.099	0/18	0/18	0/18	2/18	0/18	11%
	this work (below)	3.25	8/20	0.701	2/4	1/4	1/4	2/4	2/4	50%
	Trellis BMA (d)	4.75	6/160	0.095	1/32	1/32	0/32	4/32	0/32	12%
DNAformer nanopore	this work	2.5	0/60	0.085	0/12	0/12	0/12	0/12	0/12	0%
	this work (below)	2.25	11/20	0.819	2/4	3/4	2/4	1/4	3/4	75%
	Trellis BMA (d)	3	0/60	0.085	0/12	0/12	0/12	0/12	0/12	0%
Microsoft CNR	this work	3.5	1/80	0.089	0/16	0/16	1/16	0/16	0/16	6%
	this work (below)	3.25	10/20	0.782	3/4	1/4	2/4	2/4	2/4	75%
	Trellis BMA (d)	6	0/60	0.085	0/12	0/12	0/12	0/12	0/12	0%
Chandak exp. 0	this work	2.75	0/60	0.085	0/12	0/12	0/12	0/12	0/12	0%
	this work (below)	2.5	25/300	0.133	8/60	4/60	5/60	4/60	4/60	13%
	Trellis BMA (d)	3.25	0/60	0.085	0/12	0/12	0/12	0/12	0/12	0%
Antkowiak File 1	this work	3.5	0/60	0.085	0/12	0/12	0/12	0/12	0/12	0%
	this work (below)	3.25	15/20	0.942	3/4	4/4	3/4	3/4	2/4	100%
	Trellis BMA (d)	4.5	0/60	0.085	0/12	0/12	0/12	0/12	0/12	0%
Lietard capped	this work	2.75	0/60	0.085	0/12	0/12	0/12	0/12	0/12	0%
	this work (below)	2.5	8/20	0.701	2/4	0/4	2/4	0/4	4/4	100%
	Trellis BMA (d)	3.25	0/60	0.085	0/12	0/12	0/12	0/12	0/12	0%

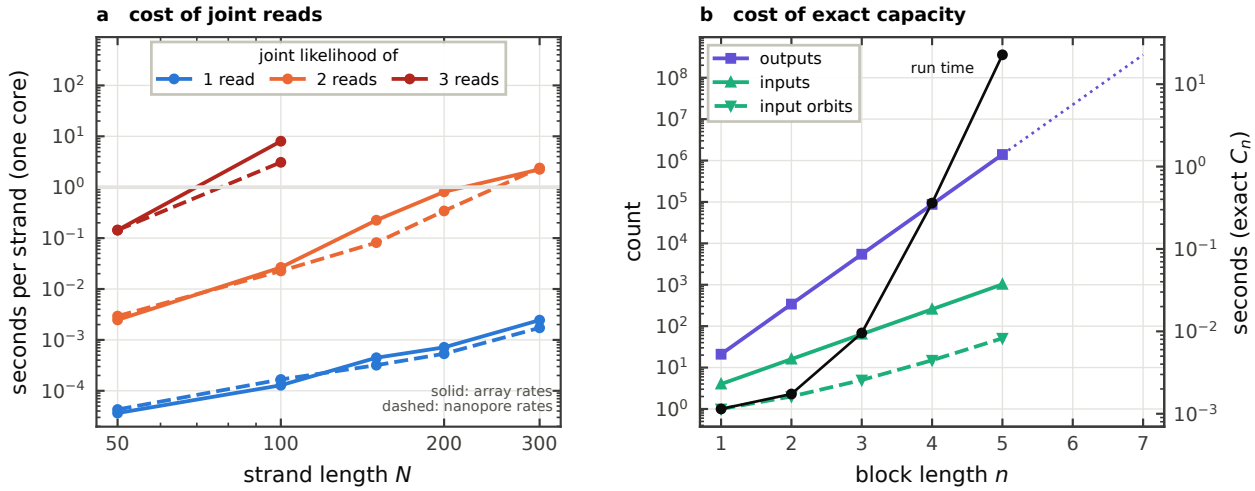
Supplementary Table 20: Prespecified confirmation run. Fixed $n = 200$ trials per point, chosen before the run from the exploratory thresholds; each trial is a fresh file (new outer code, interleaver and message, code seeds not used before) on design set $t \bmod 5$, with a fresh read subsample; at a given coverage both receivers decode the same files and reads. Failures/trials in total, exact one-sided 99% upper limit of the file error rate (valid for a fixed sample of independent files given the five design sets), and failures/trials on each design set. The design sets were drawn independently from the pool and may share designs (Methods); generalization is to new codes, messages and read samples on these designs, and the per-set counts show how much the five design sets differ.

Dataset	receiver	c	failures	99% UL	set 0	set 1	set 2	set 3	set 4
Song 0 d	this work	3.5	3/200	0.049	0/40	0/40	0/40	0/40	3/40
	Trellis BMA at c_{ours}	3.5	25/200	0.189	0/40	4/40	0/40	10/40	11/40
	Trellis BMA	3.75	1/200	0.033	0/40	0/40	0/40	1/40	0/40
DNAformer Illumina	this work	3	2/200	0.041	0/40	0/40	0/40	2/40	0/40
	Trellis BMA	3	3/200	0.049	0/40	0/40	0/40	2/40	1/40
Organick id20‡	this work	3	0/200	0.023	0/40	0/40	0/40	0/40	0/40
	Trellis BMA at c_{ours}	3	191/200	0.982	37/40	40/40	39/40	39/40	36/40
	Trellis BMA	3.5	7/200	0.078	0/40	5/40	1/40	1/40	0/40
Meiser unaged‡	this work	3.5	9/200	0.092	0/40	1/40	0/40	7/40	1/40
	Trellis BMA at c_{ours}	3.5	200/200	1.000	40/40	40/40	40/40	40/40	40/40
	Trellis BMA	4.75	8/200	0.085	0/40	3/40	0/40	4/40	1/40
DNAformer nanopore	this work	2.5	0/200	0.023	0/40	0/40	0/40	0/40	0/40
	Trellis BMA at c_{ours}	2.5	166/200	0.887	38/40	29/40	34/40	28/40	37/40
	Trellis BMA	3	0/200	0.023	0/40	0/40	0/40	0/40	0/40
Microsoft CNR	this work	3.5	3/200	0.049	0/40	0/40	1/40	2/40	0/40
	Trellis BMA at c_{ours}	3.5	200/200	1.000	40/40	40/40	40/40	40/40	40/40
	Trellis BMA	6	0/200	0.023	0/40	0/40	0/40	0/40	0/40
Chandak exp. 0	this work	2.75	0/200	0.023	0/40	0/40	0/40	0/40	0/40
	Trellis BMA at c_{ours}	2.75	200/200	1.000	40/40	40/40	40/40	40/40	40/40
	Trellis BMA	3.25	0/200	0.023	0/40	0/40	0/40	0/40	0/40
Antkowiak File 1	this work	3.5	2/200	0.041	0/40	1/40	1/40	0/40	0/40
	Trellis BMA at c_{ours}	3.5	200/200	1.000	40/40	40/40	40/40	40/40	40/40
	Trellis BMA	4.5	0/200	0.023	0/40	0/40	0/40	0/40	0/40
Lietard capped	this work	2.75	0/200	0.023	0/40	0/40	0/40	0/40	0/40
	Trellis BMA at c_{ours}	2.75	195/200	0.994	39/40	40/40	40/40	36/40	40/40
	Trellis BMA	3.25	0/200	0.023	0/40	0/40	0/40	0/40	0/40

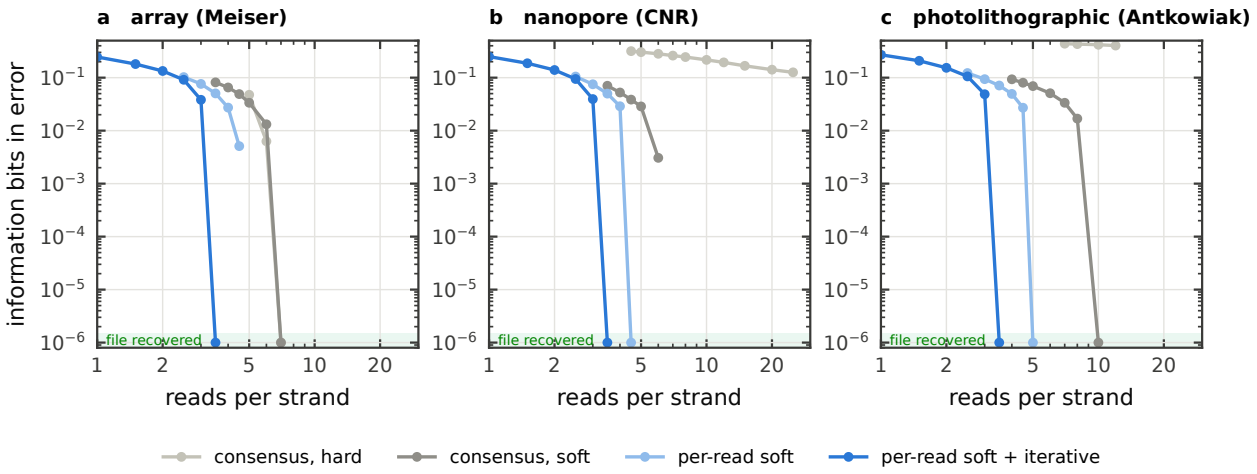
Supplementary Figures



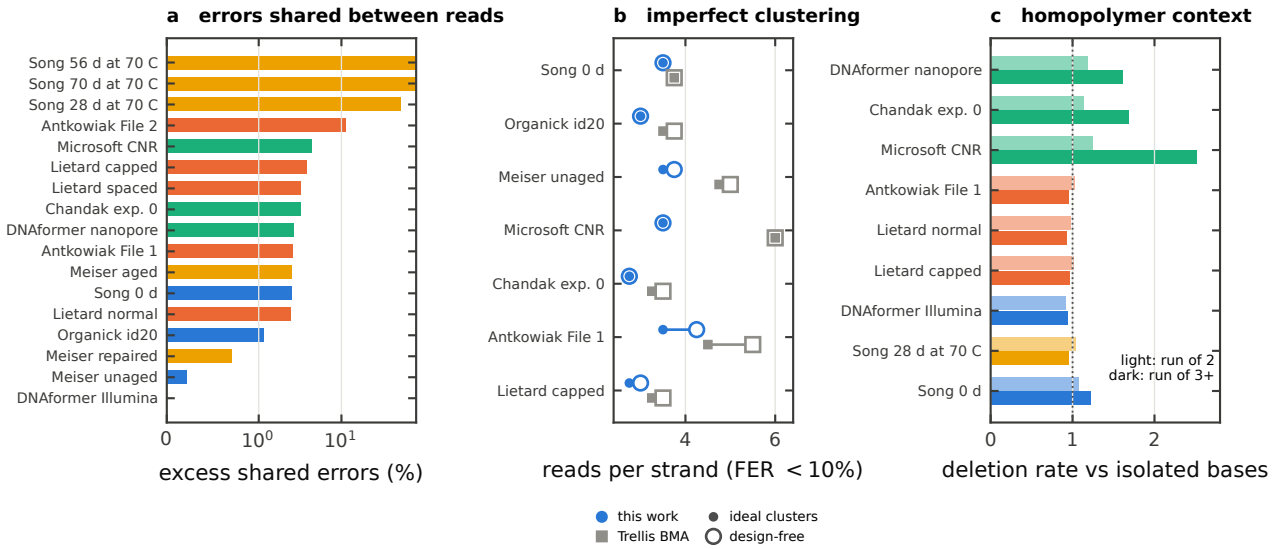
Supplementary Fig. 1: Error profiles fitted from raw reads. Per-position deletion and substitution probabilities and per-slot insertion probabilities of the fitted positional model along the payload (primers removed); colours indicate technology.



Supplementary Fig. 2: Computational cost. **a**, Time to compute the likelihood of a strand given one, two or three reads under the uniform input, versus strand length (one core; solid: array rates, dashed: nanopore rates). **b**, Size of the exact block-capacity problem (inputs, orbits of base permutations, outputs) and its run time; dotted: outputs for $n = 6, 7$.



Supplementary Fig. 3: Decoding waterfalls (historical 2-of-3 screen, not the reported result). Fraction of information bits in error after decoding versus mean coverage under the screening protocol of the earlier version (mean of three read subsamples of one strand set). These curves are a descriptive screen only and must not be quoted as the result; reported coverages are those of Table 3 for an array (**a**), a nanopore (**b**) and a photolithographic (**c**) dataset, for the iterative receiver, the per-read soft receiver without iterations, and the two consensus receivers; the shaded band marks recovery of the whole file.



Supplementary Fig. 4: Robustness. **a**, Excess probability that an error of one read recurs in another read of the same strand, beyond positional hot spots. **b**, Mean coverage at which the file error rate is below 10% with 99% confidence, with ideal clusters (small filled markers) and with design-free clustering in every trial (large open markers), for our receiver and Trellis BMA (Supplementary Table 18). **c**, Deletion rate per base in homopolymer runs of length two and of three or more, relative to isolated bases (designs of Organick and Meiser contain no runs).

References

[1] A. Kalai, M. Mitzenmacher, and M. Sudan. Tight asymptotic bounds for the deletion channel with small deletion probabilities. In *Proc. IEEE Int. Symp. Information Theory (ISIT)*, 2010.

[2] Y. Kanoria and A. Montanari. Optimal coding for the binary deletion channel with small deletion probability. *IEEE Transactions on Information Theory*, 59(10):6192–6219, 2013.

[3] R. L. Dobrushin. Shannon's theorems for channels with synchronization errors. *Problems of Information Transmission*, 3(4):11–26, 1967.

[4] D. Fertonani, T. M. Duman, and M. F. Erden. Bounds on the capacity of channels with insertions, deletions and substitutions. *IEEE Transactions on Communications*, 59(1):2–6, 2011.

[5] A. Maurer and M. Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proc. Conf. Learning Theory (COLT)*, 2009.

[6] A. Ganti, A. Lapidoth, and I. E. Telatar. Mismatched decoding revisited: general alphabets, channels with memory, and the wide-band limit. *IEEE Transactions on Information Theory*, 46(7):2315–2328, 2000.

[7] I. Shomorony and R. Heckel. DNA-based storage: Models and fundamental limits. *IEEE Transactions on Information Theory*, 67(6):3675–3689, 2021.

[8] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi. Coding over sets for DNA storage. *IEEE Transactions on Information Theory*, 66(4):2331–2351, 2020.

[9] N. Weinberger and N. Merhav. The DNA storage channel: Capacity and error probability bounds. *IEEE Transactions on Information Theory*, 68(9):5657–5700, 2022.

[10] S. R. Srinivasavaradhan, S. Gopi, H. D. Pfister, and S. Yekhanin. Trellis BMA: Coded trace reconstruction on IDS channels for DNA storage. In *Proc. IEEE Int. Symp. Information Theory (ISIT)*, pages 2453–2458, 2021.