Adaptively integrated sequencing and assembly of near-complete genomes

Hasindu Gamaarachchi[1,2] *, Igor Stevanovski[1] *, Jillian M. Hammond[1], Andre L. M. Reis[1,3], Melissa Rapadas[1], Kavindu Jayasooriya[1,2], Tonia Russell[1], Dennis Yeow[1,4-7], Yvonne Hort[1], Andrew J. Mallett[8,9,10], Elaine Stackpoole[11], Lauren Roman[12,13], Luke W. Silver[14,15], Carolyn J. Hogg[14,15], Lou Streeting[16], Ozren Bogdanovic[17,18], Renata Coelho Rodrigues Noronha[19], Luís Adriano Santos do Nascimento[19], Adauto Lima Cardoso[18-20], Arthur Georges[21], Haoyu Cheng[22], Hardip R. Patel[23], Kishore Raj Kumar[1,3,4,5], Amali C. Mallawaarachchi[1,3,24], Ira W. Deveson[1,3] #

1. Genomics and Inherited Disease Program, Garvan Institute of Medical Research, Sydney, NSW, Australia

2. School of Computer Science and Engineering, University of New South Wales, Sydney, NSW, Australia

3. Faculty of Medicine and Health, St Vincent's Healthcare Clinical Campus, University of New South Wales, Darlinghurst, NSW, Australia

4. Molecular Medicine Laboratory and Neurology Department, Concord Repatriation General Hospital, Concord, NSW, Australia

5. Faculty of Medicine and Health, University of Sydney, Camperdown, Australia.

6. Neurodegenerative Service, Prince of Wales Hospital, Randwick, Australia

7. Neuroscience Research Australia, Randwick, Australia

8. College of Medicine and Dentistry, James Cook University, Townsville, Australia

9. Department of Renal Medicine, Townsville University Hospital, Townsville, Australia

10. Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia

11. Genetic Health Western Australia, King Edward Memorial Hospital, Perth, Australia

12. Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Australia

13. CSIRO Environment, Hobart, Australia

14. School of Life and Environmental Sciences, The University of Sydney, Sydney, Australia

15. Australian Research Council Centre of Excellence for Innovations in Peptide and Protein Science, University of Sydney, Sydney, Australia

16. School of Environmental and Rural Science, University of New England, Armidale, Australia

17. School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, Australia

18. Centro Andaluz de Biología del Desarrollo, CSIC-Universidad Pablo de Olavide-Junta de Andalucía, Seville, Spain

19. Laboratório de Genética e Biologia Celular, Centro de Estudos Avançados da Biodiversidade, Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, Brazil

20. Instituto de Biociências de Botucatu, Universidade Estadual Paulista, Botucatu, Brazil

21. Institute for Applied Ecology, University of Canberra, Bruce, Australia

22. Department of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, USA

23. National Centre for Indigenous Genomics, John Curtin School of Medical Research, Australian National University, Acton, Australia

24. Clinical Genetics Service, Institute of Precision Medicine and Bioinformatics, Royal Prince Alfred Hospital, Sydney, Australia

* Contributed equally

Correspondence: i.deveson@garvan.org.au

ABSTRACT

Recent advances in long-read sequencing (LRS) and assembly algorithms have made it possible to create highly complete genome assemblies for humans, animals, plants and other eukaryotes. However, there is a need for ongoing development to improve accessibility and affordability of the required data, increase the range of usable sample types, and reliably resolve the most challenging, repetitive genome regions. 'Cornetto' is a new experimental paradigm in which the genome assembly process is adaptively integrated with programmable selective nanopore sequencing, with target regions being iteratively updated to focus LRS data production onto the unsolved regions of a nascent assembly. This improves assembly quality and streamlines the process, both for human individuals and diverse non-human vertebrates, including endemic Australian endangered species, tested here. Cornetto enables us to generate highly complete diploid human genome assemblies using only a single LRS platform, surpassing the quality of previous efforts at a fraction of the cost. Cornetto enables genome assembly from challenging sample types like human saliva, for the first time, further enhancing accessibility. Finally, we obtain complete and accurate assemblies for clinically-relevant repetitive loci at the extremes of the genome, demonstrating valid approaches for genetic diagnosis in facioscapulohumeral muscular dystrophy (FSHD) and *MUC1*-autosomal dominant tubulointerstitial kidney disease (*MUC1*-ADTKD) – inherited diseases for which diagnosis is complicated by an inability to sequence the genes involved. In summary, Cornetto will improve, accelerate and democratise genome assembly, delivering impacts across a range of bioscience domains.

INTRODUCTION

The capacity to obtain high quality and even complete telomere-to-telomere (T2T) assemblies for large eukaryotic genomes is transforming our understanding of genome architecture, variation and evolution, and will lead to improvements in genomic disease diagnosis[1,2]. The first complete T2T human genome was published in 2022, overcoming technical challenges that had left the final 8% of its sequence unsolved for two decades after the conclusion of the Human Genome Project[3]. Recent advances in the field have been driven predominantly by a handful of current US-led consortium projects, including the T2T Consortium[3], the Human Pangenome Reference Consortium (HPRC)[4] and Vertebrate Genome Project (VGP)[5]. These critical initiatives have led the way in molecular and computational methods development for eukaryotic genome assembly and evaluation[3–14].

However, concentration of research and innovation within major consortium projects also reflects the high cost and high degree of technical expertise involved in producing a complete and accurate genome assembly. Current best practices call for a combination of deep Pacific Biosciences (PacBio) 'HiFi' LRS data, coupled with Oxford Nanopore Technologies (ONT) 'ultra-long' LRS data and Illumina 'HiC' chromatin conformation capture, or an analogous 'long-range' sequencing method[1]. Integration of these different data types helps to address blindspots in each. For example, the higher accuracy of PacBio HiFi is useful for untangling segmentally duplicated DNA with fine sequence differences between copies, ONT's longer reads have capacity to span large repeats or extended regions of homozygosity, and HiC enables long-range phasing to resolve haplotypes at chromosome scale[1]. This recipe requires access to three sequencing platforms, comes with onerous sample requirements and considerable cost.

Therefore, there is a need for ongoing development to improve the affordability and accessibility of data production; increase the breadth of usable sample types and qualities; and improve data quality and assembly algorithms to better resolve the genome's most challenging regions. Failure to address these barriers will ensure the continued exclusion of many potential research projects, cohorts and species from this new era of complete genomes.

Here we present a novel genome assembly strategy designed to meet this need. ONT's 'ReadUntil' or 'adaptive sampling' functionality enables programmable selective LRS by accepting or rejecting DNA fragments, based on their sequence, in real-time[15]. This can be used to enrich genomic regions of interest, enabling targeted analysis of clinically relevant genes, for example[16–18]. We have adapted this capability to the challenge of genome assembly, integrating selective sequencing with the assembly process to enrich LRS data where it is most needed, thereby reducing production costs, sample requirements and improving assembly quality (**Fig1a**). Our new strategy is suitable for human and non-human genomes, and resolves highly repetitive medically-relevant loci and hemizygous sex chromosomes, all with exceptional accuracy.

RESULTS

Cornetto: integrated adaptive sequencing and assembly

Most of the euchromatic human genome is relatively easy to assemble using current LRS data. For example, we sequenced DNA from the hg002 reference sample on a single PacBio SMRT cell (~25x depth) and assembled the data with *hifiasm*[7]. In the resulting primary assembly, less than 10% of the genome sequence remains in contigs shorter than ~5 Mbase (**Fig1b-i**). The assembly may be improved with further whole-genome LRS, however, there is diminishing marginal utility because most of the genome is already resolved (**Fig1b-i**). Instead, we reasoned that ONT ReadUntil[15] could be used to enrich for regions of the genome that are difficult to assemble. Rather than defining these target regions within the human reference genome based on prior knowledge, a more agnostic and efficient approach is to identify solved regions within a starting assembly of moderate quality, then program these for rejection during subsequent ONT sequencing so as to enrich for unsolved regions.

This idea forms the basis for an integrated sequencing and assembly paradigm, which we nickname 'Cornetto' (see **Methods**). To establish the method, we took the hg002 primary assembly above as an initial reference, identifying short contigs (< 800 Mbase), regions adjacent to the end of a contig (within 200 kb), and regions with poor coverage, mapability or assembly quality, then labelling the remaining ~89% of the assembly as 'boring bits'. We then performed ONT duplex sequencing using the software *ReadFish*[15] to programmably reject DNA fragments originating from any of the boring bits, in real-time (**Fig1a**). We used ONT duplex data here, because it is sufficiently similar in per-base accuracy to be co-assembled with PacBio HiFi data (**Extended Data Fig1a-c**). The experiment was paused at regular intervals, allowing new and existing data to be aggregated and reassembled (**Fig1a**). The experiment was then resumed using the new assembly and an updated list of boring bits for rejection. The assembly and target selection processes are automated, with no manual curation required. The assembly was iteratively improved, expanding the boring bits and, thereby, focusing new data onto an increasingly small, unsolved fraction of genome (**Fig1a**).

Human genome assembly with PacBio and ONT data

We performed two independent experiments with DNA from hg002 (*hg002-Cornetto-1* and *hg002-Cornetto-2*). Each was sequenced with one PacBio SMRT cell and three ONT duplex flow cells, which were run in succession according to the iterative Cornetto process above (3x cycles per flow cell). Comparing the primary assemblies, we observed incremental improvements in contiguity over the course of the experiments, resulting in substantial overall gains (**Fig1b-iii,iv**). For example, *hg002-Cornetto-1* was improved from 496 contigs with an N_{50} length of 54.5 Mbase and N_{90} of 6.3 Mbase to a final assembly of 130 contigs with N_{50} of 134.1 Mbase (2.5-fold improvement) and N_{90} of 50.4 Mbase (8.0-fold improvement; **Fig1b-iii,iv**; **Extended Data Table 1**). Whereas no chromosome was assembled as a single primary contig in the initial assemblies, we obtained 15 in the final assembly for *hg002-Cornetto-1* and 11 for *hg002-Cornetto-2* (**Fig1c**).

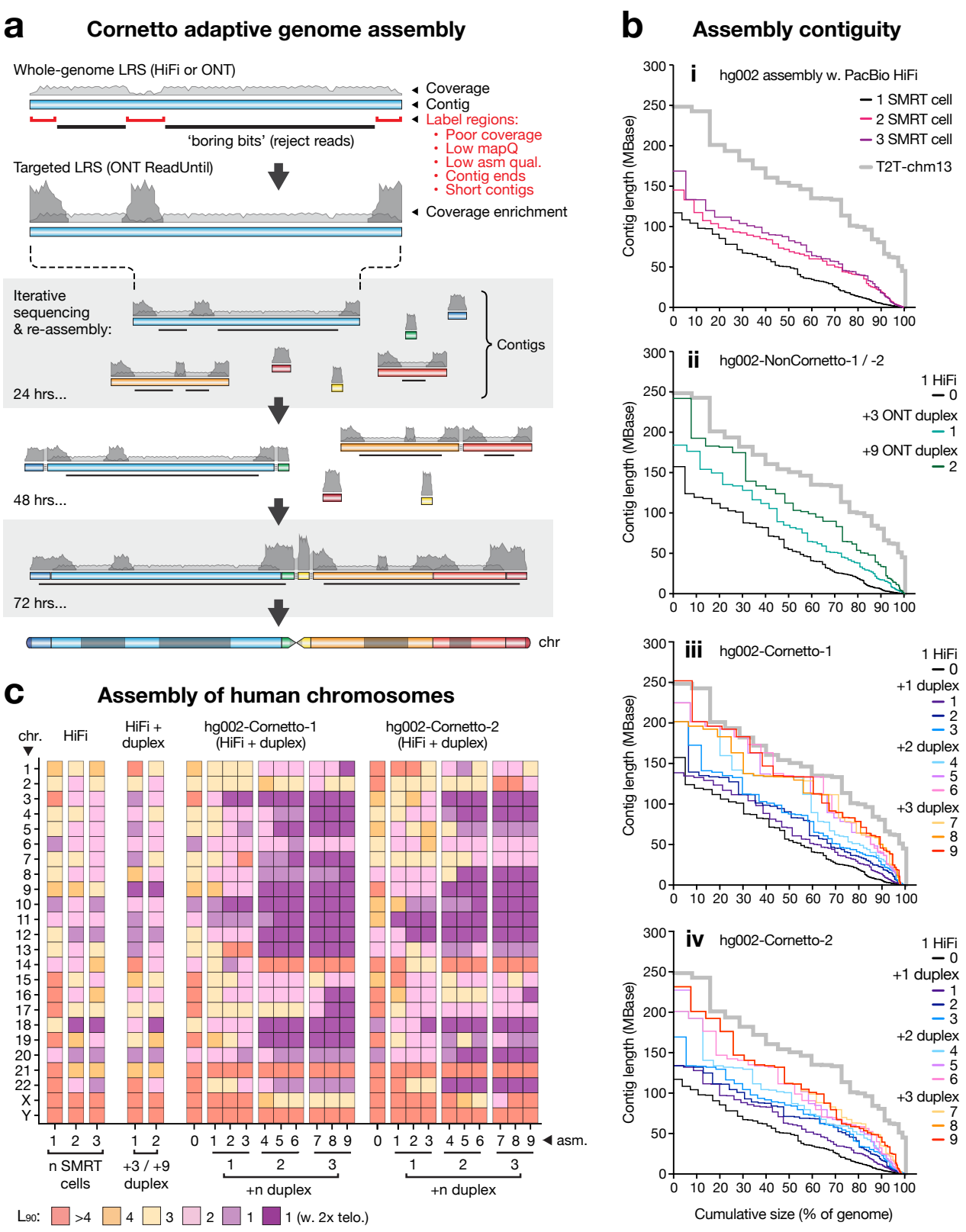

Figure 1. Human genome primary assemblies. (**a**) Overview of the Cornetto method. A primary assembly generated from whole-genome LRS data (PacBio HiFi or ONT), provides the starting reference. Regions with poor coverage, mappability, assembly quality, short contigs and contig ends are identified. The remainder are considered 'boring bits' and labelled for rejection by ONT ReadUntil. Sequencing is paused at regular intervals (e.g. 24, 48, 72hrs) and new assembly is generated, providing an updated reference and boring bits. Sequencing is resumed after washing and re-loading flow cell. Data is focused onto an increasingly small and challenging portion of the genome. (**b**) For a given primary assembly, Nx plots show contigs sizes sorted from largest to smallest, relative to cumulative assembly size, as a percentage of human genome (3.1 Gbase). Assemblies were generated using LRS data from hg002. T2T-chm13 is shown for comparison (grey line). Sub-panels show: (**i**) Assemblies of HiFi data from 1, 2 or 3 SMRT cells. (**ii**) Combined data from 1 SMRT cell and 3 ONT duplex flow cells (hg002-NonCornetto-1); or 1 SMRT cell and 9 duplex flow cells (hg002-NonCornetto-2). (**iii, iv**) Two Cornetto experiments (hg002-Cornetto-1 / -2), each using 1 SMRT Cell and 3 duplex flow cells run sequentially with 3 cycles per cell, resulting in 9 intermediate assemblies. (**c**) For the same assemblies, tile plot shows contiguity of human chromosomes. Colour scale encodes L90 values: number of contigs encompassing >90% of the reference sequence for a given chromosome. Dark purple tiles show chromosomes with L90 = 1 and a telomere detected at each end, indicating the whole chromosome is assembled as a single primary contig.

To put these results in context, we generated a matched assembly from the same PacBio HiFi data, this time augmented with three ONT duplex flow cells run without adaptive sequencing (*hg002-NonCornetto-1*; **Extended Data Fig1a-c**). The initial assembly was improved, but the gains were small compared to those achieved using Cornetto. For example, N_{50} and N_{90} lengths were improved by 1.5-fold and 2.7-fold, respectively, in *hg002-NonCornetto-1* compared to 2.5-fold and 8-fold for *hg002-Cornetto-1* (**Fig1b-ii,c**). We further augmented the non-Cornetto assembly by adding published ONT duplex data[8] (*hg002-NonCornetto-2*). However, even with up to nine duplex flow cells – beyond which there were no further gains – we were unable to obtain a primary assembly of comparable contiguity to *hg002-Cornetto-1* or *-2* (**Fig1b-ii,c**). Assemblies generated using Cornetto were also equivalent or superior across a range of standard quality metrics, including per-base accuracy (QV), BUSCO gene completeness, and rates of duplicated or fragmented genes (**Extended Data Table 1**). These results establish the capacity of our Cornetto strategy to efficiently harness LRS data for improved assembly of human genomes.

ONT-only diploid human genome assemblies

To further streamline the assembly process, we next tested the Cornetto paradigm using ONT data alone. We generated a new hg002 assembly (*hg002-Cornetto-3*), this time with data from a standard ONT flow cell (LSK114; simplex reads; **Extended Data Fig1a-c**) augmented with a second flow cell run with Cornetto adaptive sequencing (3x cycles). As above, the primary assembly was improved via Cornetto, obtaining a final N_{50} of 154.4 Mbase (1.6-fold improvement) and N_{90} of 79.1 Mbase (2.1-fold improvement; **Extended Data Fig2a-d**; **Extended Data Table 2**).

Given this promising result, we adapted the Cornetto paradigm toward the challenge of producing diploid genome assemblies, noting that all results reported so far refer to primary assemblies (i.e. where maternal and paternal haplotypes remain collapsed into a single, linear representation). To do so, phased contigs produced by *hifiasm* during each Cornetto cycle were aligned to their corresponding primary assembly, to identify regions not spanned by contigs from both haplotypes (**Fig2a**). These unphased regions were excluded from the list of boring bits, which were otherwise defined as above (see **Methods**). We reasoned that the enrichment of coverage in these regions may be beneficial for closing gaps in phasing – by providing additional reads that may span a homozygous region, for example – thereby improving the resulting diploid assembly (**Fig2a**).

We used our modified Cornetto strategy to generate a diploid hg002 assembly (*hg002-Cornetto-4*), again using one standard ONT flow cell and a second run with Cornetto adaptive sequencing. We obtained a highly complete diploid assembly, with haplotypes exhibiting N_{50} lengths of 132.2 and 135.7 Mbase (2.6-fold improvement) and N_{90} lengths of 35.6 and 61.3 Mbase (8.1-fold improvement; **Fig2b**; **Extended Data Fig3a**). *Hg002-Cornetto-4* contained 27 complete chromosomes out of a possible 46, compared to just 3/46 prior to Cornetto (**Fig2c,d**; **Extended Data Fig3b**). This included complete pairs for ten autosomes and, notably, both chrX and chrY were fully assembled despite their repetitive architectures (**Fig2d**). Alignment of *hg002-Cornetto-4* to the Q100 project *T2T-hg002* reference, taken here as a ground truth, confirmed the assembly is highly complete, accurate and free of large misassemblies (**Fig2c**; **Extended Data Fig3c**).

A recent T2T Consortium study presented a strategy for assembling diploid human genomes using data from ONT instruments alone, doing so with a combination of 50x duplex, 30x ultra-long and 50x pore-C data, utilising >15 ONT flow cells in total[8]. We evaluated our *hg002-Cornetto-4* assembly, which was created using a single ONT ligation library prep and just two flow cells, by comparison to this published assembly (*T2TC-ONT-only*). *Hg002-Cornetto-4* and *T2TC-ONT-only* showed similar contiguity and contained 27 vs 26 complete chromosomes, respectively (**Fig2b,d**; **Extended Data Table 2**). *Hg002-Cornetto-4* was somewhat more complete and accurate than *T2TC-ONT-only* (BUSCO 99.0% vs 98.1%; QV 56 vs 53), and switch error rates were equivalent (1.2%; **Fig2e**). The use of pore-C data (analogous to HiC) for chromosome scaffolding and phasing means that only 13 of the 26 complete chromosomes in *T2TC-ONT-only* are free of gaps, whereas *hg002-Cornetto-4*

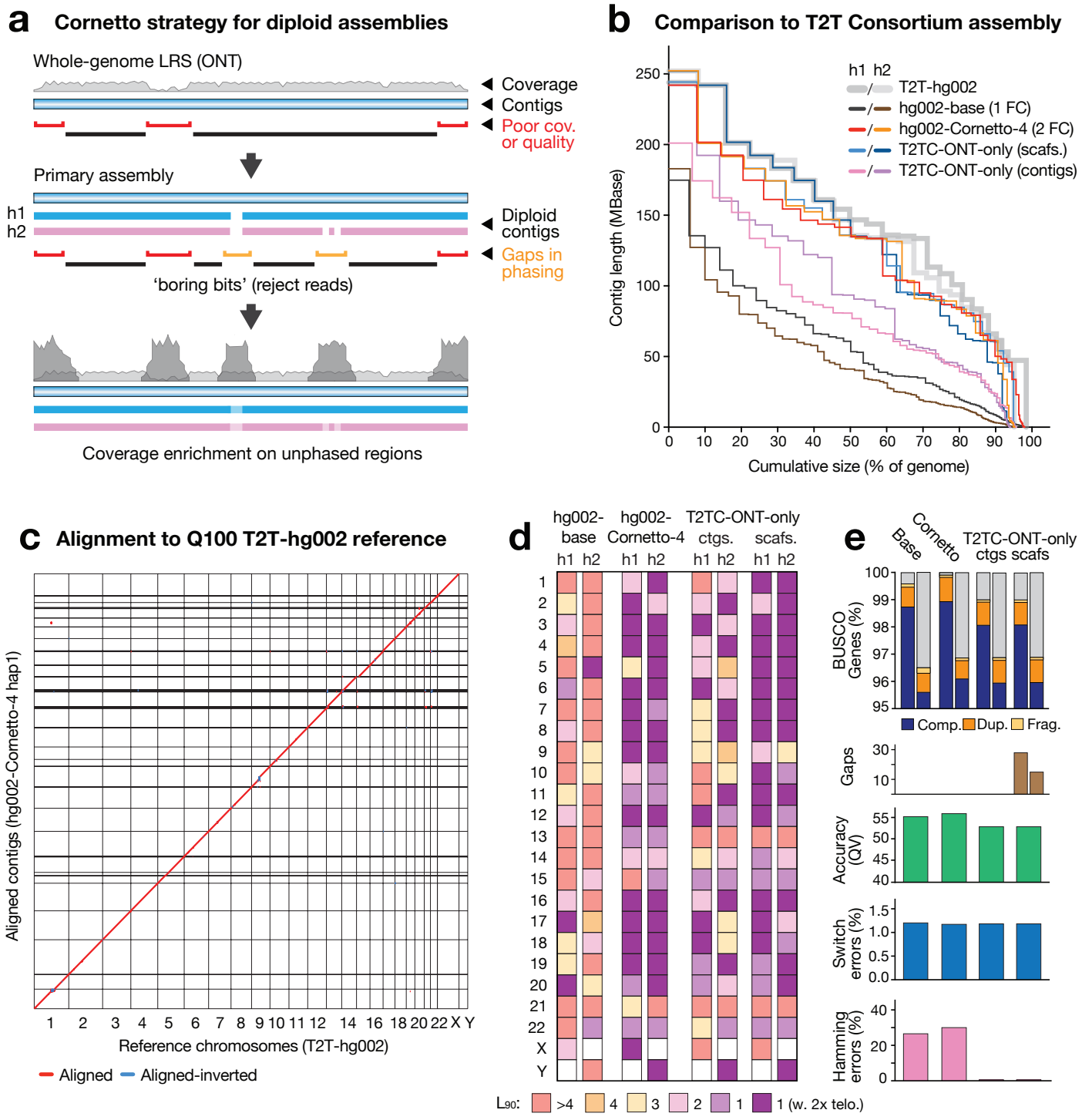

Figure 2. Nanopore-only diploid human genome assemblies. (**a**) Overview of the updated Cornetto method for improved diploid assemblies. Same process as shown in Figure 1 is followed with the additional step of excluding unphased regions from the list of boring bits. These are determined by aligning all contigs from haplotype 1 & 2 to their primary assembly and identifying any region not spanned by a contig on both haplotypes. Enrichment of ONT data in these regions may help to close phase gaps. (**b**) Nx plot shows contig/scaffold sizes sorted from largest to smallest, relative to cumulative assembly size, as a percentage of human genome (3.1 Gbase). The plot compares hg002 diploid assemblies generated with haplotypes plotted separately (h1/h2). Assemblies generated using ONT data from one flow cell (hg002-base) then augmented with a second flow cell run with Cornetto (hg002-Cornetto-4) are compared to an ONT-only assembly from the T2T Consortium (T2TC-ONT-only), plotted as scaffolds vs contigs. The Q100 T2T-hg002 assembly provides a reference. (**c**) Dot plot shows alignment of hg002-Cornetto-4 contigs (vertical axis) to chromosomes in Q100 T2T-hg002 (horizontal axis). The plot shows haplotype 1 and haplotype 2 is in Extended Data Fig3. (**d**) Tile plot shows contiguity of human chromosomes in same assemblies as b. Colour scale encodes L_{90} values: number of contigs encompassing >90% of the reference sequence for a given chromosome. Dark purple tiles show chromosomes with L_{90} = 1 and a telomere detected at each contig end, indicating the whole chromosome is assembled into a single contig. (**e**) Bar charts compare assembly quality metrics: proportion of BUSCO genes detected as complete, duplicated, fragmented or missing; total number of gaps; sequence accuracy, as per QV values (k-mer size of 21); switch errors (%); hamming errors (%).

contains no gaps (**Fig2d,e**). Conversely, chromosomes in *hg002-Cornetto-4* are only partially phased, reflected in its high rate of hamming errors (30.2%; **Fig2e**). We noted this could be addressed by using parental sequencing data for long-range phasing of the *hg002-Cornetto-4* (**Extended Data Table 2**). Overall, while not a perfect like-for-like comparison, our ONT-only diploid human assembly *hg002-Cornetto-4* is equivalent or superior to *T2TC-ONT-only* on most metrics, despite being created with a fraction of the resources.

Assembling medically-relevant repetitive loci

The human genome contains hundreds of analytically challenging repetitive loci with known roles in disease[12]. To illustrate the potential for Cornetto to improve inherited disease diagnosis, we next explored two such loci, which are among the most extreme known examples. In both cases, a current inability to sequence the causative locus is a barrier to effective diagnosis for its relevant disease, namely facioscapulohumeral muscular dystrophy (FSHD) and *MUC1*-autosomal dominant tubulointerstitial kidney disease (*MUC1*-ADTKD). The unmet needs and challenges involved are described in more detail in **Supplementary Notes 1** and **2**, respectively.

FSHD is a progressive myopathy resulting from aberrant expression of the *DUX4* gene residing within the 4q D4Z4 macrosatellite repeat, a polymorphic *n* x 3.3kb tandem repeat in the sub-telomeric region of chr4q[19]. The repeat typically ranges in size from ~11–100 copies[19]. FSHD most commonly presents in individuals with a contracted 4q D4Z4 haplotype (<10 copies), which must also harbour a permissive sequence variant (4qA) following the distal-most *DUX4* copy[19]. To assess our capacity to accurately assemble this locus, we extracted the sub-telomeric region from both copies of chr4q in the *hg002-Cornetto-4* assembly, annotated known sequence features relevant to FSHD, then compared them to the equivalent regions of the *T2T-hg002* reference, again taken as ground truth. In both assemblies we identified one D4Z4 haplotype at 42 copies in length (~139 kb) of subtype 4qA and a second at 26 copies (~86 kb) of subtype 4qB (**Fig3a**). Aligning corresponding haplotypes between the two assemblies, we observed 99.99% and 99.97% sequence concordance across entire 4q D4Z4 regions (**Fig3a**). We next performed targeted ONT sequencing and assembly of this region in four patients with diagnostically confirmed FSHD. In each case we identified one D4Z4 haplotype of the permissive 4qA sub-type with fewer than 10 copies, sufficient for a positive diagnosis, and observed repeat lengths that were concordant with previous molecular genetic testing (**Fig3b**; see **Supplementary Note 1**).

ADTKD is a chronic kidney disease typically caused by variants in one of four genes, *UMOD*, *MUC1*, *REN* and *HNF1B*[20]. *MUC1* is thought to account for around ~20% of cases, however, diagnosis of *MUC1*-ADTKD is obscured by technical challenges in resolving this gene. *MUC1* contains a *n* x 60bp variable number tandem repeat region (VNTR)[21]. This is highly polymorphic, varying in length (20–125 copies per haplotype) and differing in the composition of imperfect sequence subunits within and between individuals[4,21]. Duplication of a cytosine (dupC) within this VNTR, which results in a frameshifting variant, has been identified as the predominant cause of *MUC1*-ADTKD[22] (**Fig4a**). We identified both copies of the *MUC1* locus within our *hg002-Cornetto-4* assembly, annotated the VNTR region for known and novel 60bp subunits, and compared them to *T2T-hg002* (**Fig4b**). In both assemblies we identified one VNTR haplotype with 65 x 60bp copies (~3.9 kb) and one with 78 x 60bp copies (~4.7 kb). The composition and order of VNTR subunits was matched and, aligning corresponding haplotypes between assemblies, we observed perfect sequence concordance across the entire VNTR region (**Fig4b**). We next performed targeted ONT sequencing and *MUC1* assembly in ten patients with diagnostically confirmed *MUC1*-ADTKD. Across 11 individuals (including *hg002-Cornetto-4*), we observed 20 unique VNTR haplotypes which ranged in size from 40–83 copies, with no individuals sharing the same pair of haplotypes (**Fig4b**). In each patient (but not hg002) we identified a single dupC frameshift variant within the VNTR occurring on a single haplotype, sufficient for a positive diagnosis. Notably, 8/10 pathogenic haplotypes were unique, implying frequent independent origins of the dupC variant (**Fig4b**; see **Supplementary Note 2**). Overall, these results establish the capacity to assemble both the 4q D4Z4 and *MUC1* loci with exceptional accuracy, providing viable new avenues to improve the genetic diagnosis of FSHD and ADTKD.

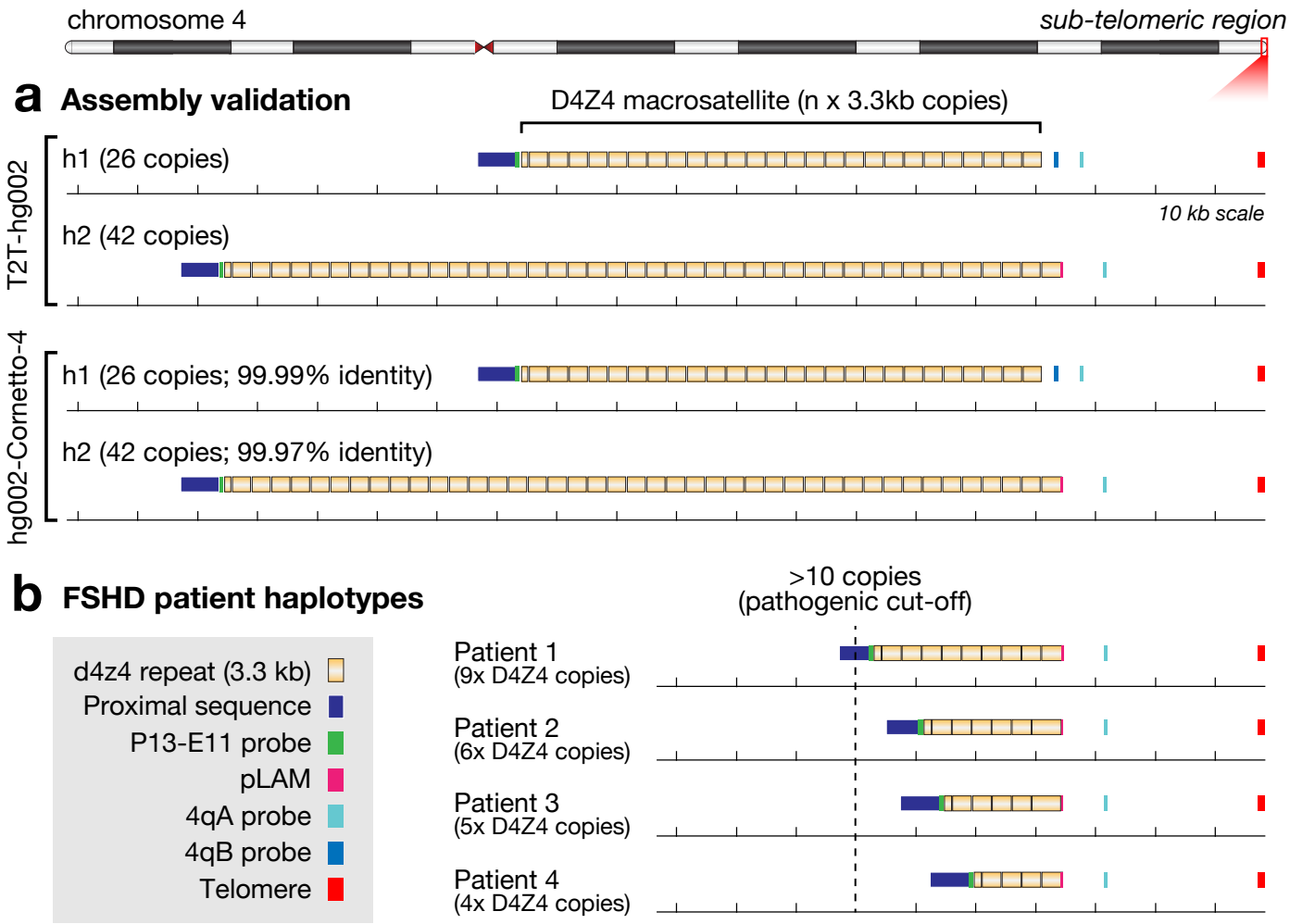

Figure 3. Assembly and genotyping of 4q D4Z4 for genetic diagnosis of FSHD. (**a**) Genome browser views show annotated sequence features within the 4q subtelomeric region on each haplotype (h1 / h2) for the Q100 T2T-hg002 reference assembly (upper) and the hg002-Cornetto-4 assembly (lower). The D4Z4 macrosatellite repeat is annotated with recurring 3.3 kb subunits in yellow. A range of other sequence features relevant for 4q D4Z4 genotyping are shown, including markers for the permissive (4qA) and non-permissive (4qB) distal DUX4 sequence variants (see **Supplementary Note 1**). The 4q D4Z4 length is stated above each haplotype. Identity scores stated for hg002-Cornetto-4 were determined by aligning the entire 4q D4Z4 sequence to the corresponding haplotype in the T2T-hg002 reference, taken as a ground truth. (**b**) Same plots as above, this time showing pathogenic haplotypes assembled using targeted ONT sequencing of the 4q D4Z4 region for four patients with diagnostically confirmed FSHD. In each case, the 4q D4Z4 length is shorter than 11 copies and harbours the permissive 4qA sequence variant, sufficient for a positive genetic diagnosis. Expected repeat sizes from previous genetic testing are stated for each patient (see **Supplementary Note 1**).

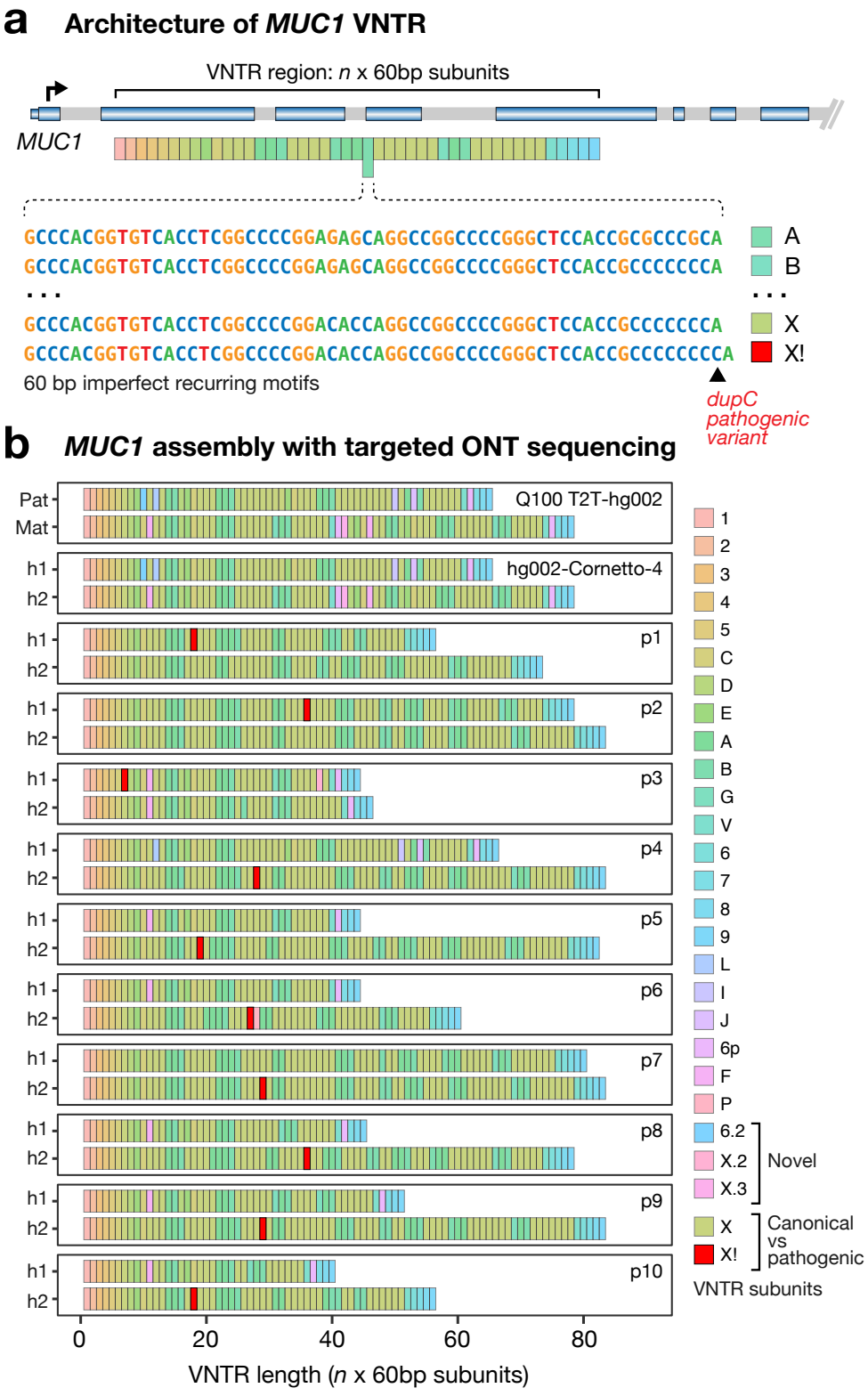

Figure 4. Assembly and genotyping of MUC1 for genetic diagnosis of ADTKD. (**a**) Schematic of MUC1 variable number tandem repeat region (VNTR) and known genetic basis for MUC1-ADTKD. Briefly, MUC1 contains a large VNTR comprising recurring imperfect 60bp subunits, which varies in length and sequence composition within and between individuals. Duplication of a cytosine (dupC) within the VNTR, resulting in a frame-shift causes MUC1-ADTKD (see **Supplementary Note 2**). (**b**) Tile-bar plots show the VNTR length and subunit composition identified for each MUC1 haplotype (h1 / h2) for the Q100 T2T-hg002 reference (upper) and hg002-Cornetto-4 assembly, which show identical length and sequence composition between corresponding haplotypes. Below are VNTR haplotypes assembled via targeted ONT sequencing in ten patients with diagnostically confirmed MUC1-ADTKD. Different coloured tiles indicate known and novel 60bp sequence subunits, with the known dupC pathogenic subunit (X!) shown in red. A single X! subunit was identified on one haplotype in each patient, sufficient for a positive diagnosis (see **Supplementary Note 2**).

Genome assemblies from human saliva

Another benefit of Cornetto is to enable production of high quality assemblies from challenging and/or limited sample types. Human saliva is one such sample type where there would be significant utility. Saliva is more easily accessible than blood or other human tissues and can be collected, shipped and stored at room temperature. This can be advantageous in some clinical contexts, for field studies in remote communities[23] or even for direct-to-consumer genomics[24]. However, saliva is less amenable than blood to extraction of high-molecular weight (HMW) DNA; is not compatible with ONT's ultra-long protocol, nor HiC or other related long-range methods; and often suffers from relatively high levels of non-human DNA contamination[25]. Given these challenges, we are not aware of previous attempts to assemble a human genome from saliva.

We collected saliva from a male and female participant, extracted HMW DNA, then conducted Cornetto sequencing and assembly on each. We tested a combined PacBio HiFi and ONT duplex approach (*saliva-A-Cornetto-1*; *saliva-B-Cornetto-1*) and an ONT-only approach (*saliva-A-Cornetto-2*; *saliva-B-Cornetto-2*). Because non-human DNA was present (**Fig5a**), we additionally included non-human contigs identified in the initial assemblies in the target list for rejection, selecting against further contamination (see **Methods**). Both Cornetto approaches yielded high-quality genome assemblies with improved contiguity and completeness relative to their starting assemblies (**Fig5b-d**; **Extended Data Fig4a-c**; **Extended Data Table 3**). The improvements were particularly pronounced for ONT-only diploid assemblies *saliva-A-Cornetto-2* and *saliva-B-Cornetto-2*, for which we obtained final contig N_{90} lengths of 46.5 Mbase (15-fold) and 50.1 Mbase (27-fold), and 27/46 and 26/46 complete chromosomes, respectively (**Fig5b,c**). Despite the use of saliva as input material, these results are comparable to the *hg002-Cornetto-4* and *T2TC-ONT-only* assemblies above.

For further context, we compared our saliva assemblies to 47 assemblies released in the first phase of the HPRC, which were generated using cultured cells and a combination of LRS and long-range techniques[4]. *Saliva-A-Cornetto-2* and *Saliva-B-Cornetto-2* exhibited comparable or superior BUSCO gene completeness and substantially better contiguity than any available HPRC assembly (**Fig5b,d**). Although assembly accuracy cannot be directly measured, as no ground truth is available, the results presented above for *hg002-Cornetto-4* imply comparability with HPRC assemblies on these parameters. In summary, Cornetto can be used to obtain highly complete assemblies from human saliva, which are in line with (or surpass) quality standards at the leading edge of the genomics field.

Genome assemblies for non-human vertebrates

Cornetto is sequence-agnostic and does not rely on any prior knowledge of the genome being assembled. In theory, the method is suitable for any species. To establish this, we next assembled genomes for a selection of non-human vertebrates from diverse lineages, prioritised for their salience in research and conservation. The critically endangered orange-bellied parrot (*Neophema chrysogaster*) and endangered western saw-shelled turtle (*Myuchelys bellii*) were assembled using only ONT data, while Gould's petrel (*Pterodroma leucoptera*; a threatened seabird) and the redstriped eartheater cichlid (*Geophagus surinamensis*; an Amazonian fish) were assembled with PacBio HiFi and ONT duplex data (see **Supplementary Note 3**).

For each species, Cornetto delivered improvements in genome assembly outcomes, compared to base assemblies generated with standard LRS data (**Fig6a-c**; **Extended Data Fig5**). For example, we obtained a 3.9-fold increase in contig length N_{50} for the petrel primary assembly and an increase in the number of chromosomes assembled as single primary contigs from 6 to 27 (**Fig6a-c**). ONT-only diploid assemblies were also strongly improved. In the turtle genome, for example, each haplotype harboured ten complete chromosomes, including examples of both macro and microchromosomes, and complete single copies for 99.8% and 99.6% of BUSCO genes (**Fig6b,c**). Importantly, these improvements were obtained despite wide variation in genome sizes, architecture, depth of starting LRS data and initial assembly quality (see **Supplementary Note 3**). For example,

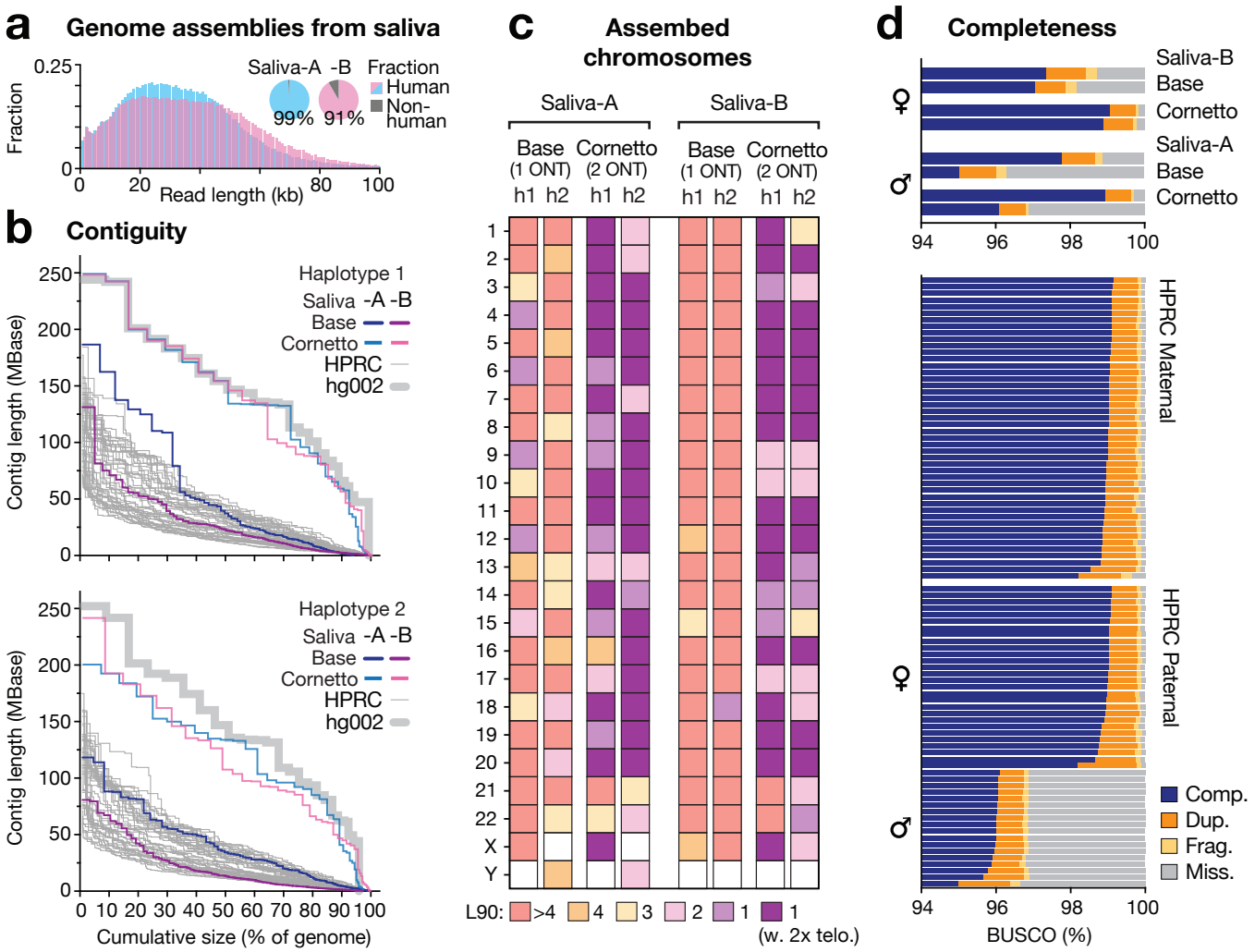

Figure 5. Genome assemblies from human saliva. (**a**) Histograms show read length profiles and pie charts show proportion of non-human reads from standard ONT sequencing on saliva samples from two participants: Saliva-A (male) and Saliva-B (female). (**b**) Nx plot shows contig sizes sorted from largest to smallest, relative to cumulative assembly size, as a percentage of the human genome size (3.1 Gbase). For each participant, assembly generated using ONT data from one flow cell (base) then augmented with a second flow cell run with Cornetto are shown (Saliva-A-Cornetto-2, Saliva-B-Cornetto-2). Non-human reads were excluded prior to assembly. For comparison, thin grey lines show contig sizes for all assemblies released in the first phase of the HPRC (n = 47) and thick grey lines show the Q100 T2T-hg002 assembly. Diploid haplotypes for each assembly are divided between the two plots. (**c**) For the same assemblies, tile plot shows contiguity of human chromosomes. Colour scale encodes L_{90} values: number of contigs encompassing >90% of the reference sequence for a given chromosome. Dark purple tiles show chromosomes with L_{90} = 1 and a telomere detected at each contig end, indicating the whole chromosome is assembled as a single contig. (**d**) For the same assemblies as b, stacked bar charts show the proportion of BUSCO genes detected as complete, duplicated, fragmented or missing. Haplotype groups containing the Y-chromosome sequence have a larger proportion of missing genes (designated with male marker symbol).

the cichlid genome was initially sequenced to 78x depth with HiFi data, yielding a base assembly of 142 contigs. Cornetto still found room for improvement, reducing the number of primary contigs to 75, with a 3.8-fold improvement in contig length N_{90} and 12 chromosomes added (**Fig6a**; **Extended Data Fig5a**). In contrast, the parrot genome was sequenced to 35x depth with standard ONT data on DNA extracted from a frozen liver sample, yielding a base assembly of >2000 contigs. With a single additional ONT flow cell run with Cornetto we were able to obtain a diploid assembly with a 4.5-fold increase in contig length N_{50} and a 4.6% increase in BUSCO completeness (92.1% vs 96.7%; **Fig6a,c**; **Extended Data Fig5n**). Notably, our updated assembly for the orange bellied parrot contained an 80kb region with three identifiable Major Histocompatibility Complex genes (*MHCI*, *II-A*, *II-B*), which are of critical importance for understanding the decline of immunogenetic diversity that threatens the survival of the species, whereas a recent assembly created with HiFi and HiC data was unable to resolve the MHC region[26] (see **Supplementary Note 3**). In summary, this establishes the suitability of our Cornetto assembly paradigm for assembling diverse non-human genomes.

DISCUSSION

Cornetto is a novel approach to genome assembly, applicable to both human and non-human genomes. We have used Cornetto to obtain highly complete, diploid human genome assemblies for hg002 (using cultured cells) and from saliva samples. These are notable both for their completeness, accuracy and the modest resources used in their creation. Our best assemblies used data from a single ONT library sequenced on two flow cells on a portable 'P2 Solo' device (roughly the size of a brick), without the need for PacBio and Illumina data, which require large instruments with substantial capital costs. Similarly, we did not use ONT ultra-long or pore-C methods. These are sensitive preparations typically requiring access to cultured cells or large volumes of freshly drawn blood. Although we show that Cornetto is compatible with ONT's highly accurate duplex data type, this was not needed to produce our best assemblies, nor were computationally expensive error correction methods using deep learning (*HERRO*[27] and *Dorado Correct*) – although these could foreseeably be used to further improve on Cornetto assemblies. Overall, Cornetto improves genome assembly outcomes, while streamlining the process and enhancing accessibility.

Although we obtained high quality genomes, the best Cornetto assembly (*hg002-Cornetto-4*) lacked complete contigs for 19 out of 46 human chromosomes. None of the acrocentric chromosomes (chr13, chr14, chr15, chr21, chr22) were fully assembled. These are characterised by repetitive ribosomal DNA arrays, which are largely intractable for current assembly algorithms[1,3]. Centromere regions were similarly problematic, as these regions have typically been tackled using ONT ultra-long reads previously[28]. Recent improvements to the *hifiasm*[7] software have been critical to the viability of the Cornetto paradigm and we anticipate future updates may help to resolve these remaining genome regions. Another limitation is the lack of full-length chromosome phasing, given HiC/pore-C was not used[13]. Where accessible, trio sequencing data can be used to address this. We also note that both ONT ultra-long and pore-C preparations are compatible with ONT selective sequencing, so may be successfully integrated with Cornetto in the future. Another intended improvement is to enable iterative updating of the genome assembly and its associated 'boring bits' in real-time during ONT sequencing. Currently, re-assembly is performed at experimental pause-points, requiring around ~4-6 hours. Significant software acceleration is needed to enable real-time assembly.

Cornetto works by selectively enriching LRS data onto unsolved regions of a nascent assembly. An alternative approach would be to select static, predefined target regions within the human reference genome, which are known to be challenging. However, this requires a high quality existing reference genome and prior knowledge to define target regions and is therefore unsuitable for most non-human species. The optimal target space may also differ between individuals based on their specific genome architectures, being strongly influenced by features such as repeat lengths and homozygous regions, which vary between individuals. The optimal target space may also differ depending on the nature of available data (read length, depth, accuracy). We therefore

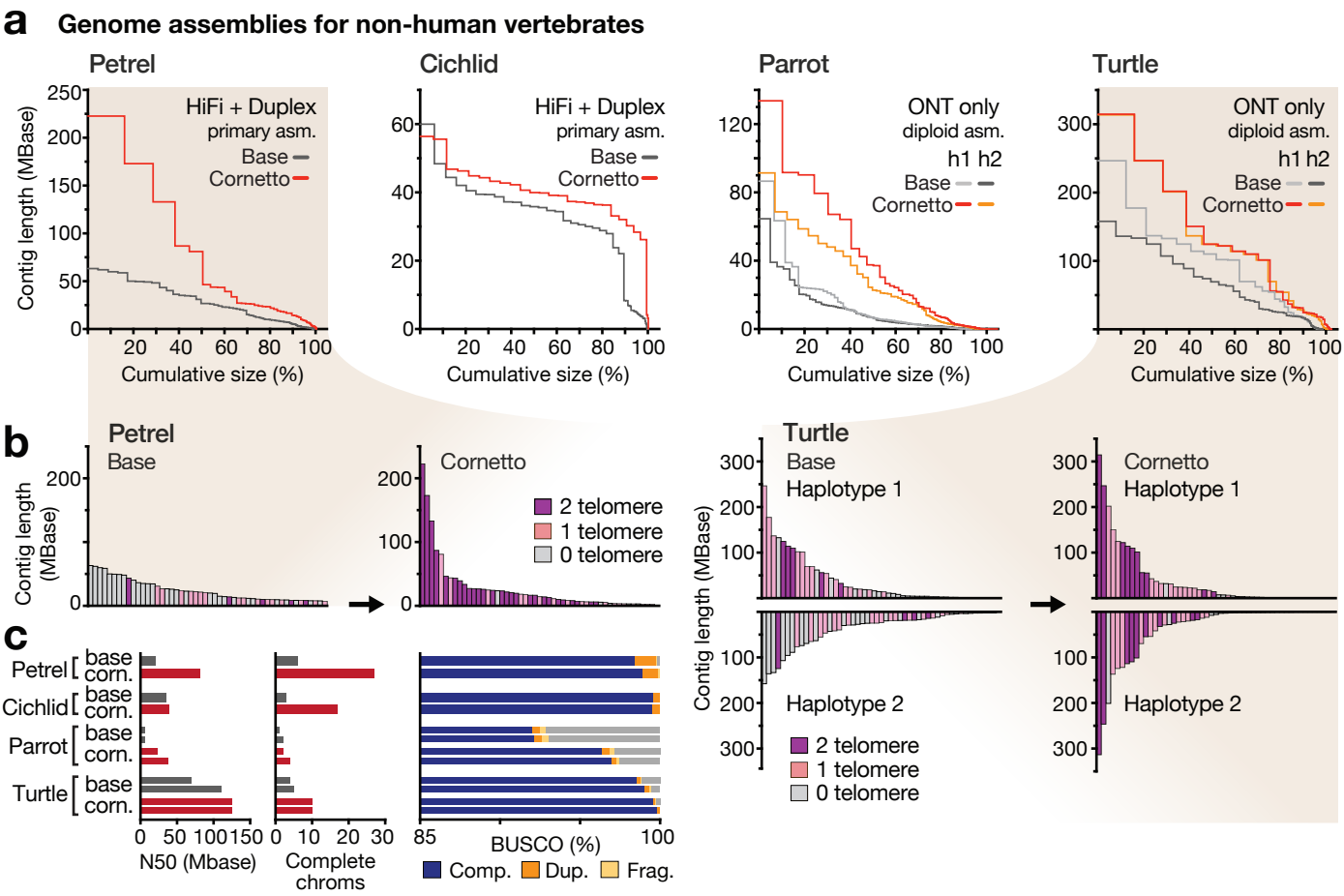

Figure 6. Genome assemblies for non-human vertebrates. (**a**) Nx plot shows contig sizes sorted from largest to smallest, relative to cumulative assembly size, as a percentage of the haploid genome size for each species. From left to right the plots show genome assemblies for: Gould's petrel (Pterodroma leucoptera); redstriped eartheater cichlid (Geophagus surinamensis); orange-bellied parrot (Neophema chrysogaster); western saw-shelled turtle (Myuchelys bellii; see **Supplementary Note 3**). The petrel and cichlid were assembled with PacBio HiFi (1 SMRT cell; base) plus ONT duplex data (2 duplex cells; cornetto). The parrot and turtle were assembled using ONT data, one or two standard flow cells (base) plus another with adaptive sequencing (Cornetto; simplex reads; LSK114). For ONT-only experiments, diploid assemblies were generated and haplotypes are plotted separately. (**b**) For the petrel and turtle assemblies above, bar plots show sizes of the fifty largest contigs in descending order, coloured according to presence of telomere sequences at contigs ends (both ends = purple; one end = pink). Equivalent plots for the cichlid and parrot are shown in **Extended Data Fig5**. (**c**) Bar plots show standard quality metrics for the same assemblies as in **a**. From left to right, these are: contig N50 lengths in Mbases; number of complete chromosomes (>1 Mbase and telomere detected at both ends); proportion of BUSCO genes detected as complete, duplicated, fragmented or missing.

believe the most efficient approach is simply to select for any region that has not yet been confidently assembled, defining this in the context of a specific genome and all available data.

However, there are scenarios where targeting static reference regions may be preferable. For example, we used this approach to specifically enrich 4q D4Z4 and *MUC1* to efficiently and reproducibly assemble these loci in patient samples. An inability to effectively sequence these loci is a major barrier to genetic diagnosis of FSHD and *MUC1*-ADTKD, respectively (see **Supplementary Notes 1** & **2**). There is also much that remains to be learned about these diseases. For example, we show here that the dupC variant known to cause *MUC1*-ADTKD has reproducibly, but independently, emerged within different families or ancestry groups, reflected in the lack of shared pathogenic haplotypes between patients. Although dupC is overwhelmingly the most common causative variant reported in *MUC1*-ADTKD[29], this is heavily biased by the available diagnostic technique being targeted to this specific variant[21]. We anticipate that the accurate, haplotype-resolved assembly of the *MUC1* VNTR in currently undiagnosed patients will reveal additional pathogenic variants.

There is clear clinical utility in being able to genotype these and other challenging medically-relevant loci from patient saliva, but developing the capacity to assemble highly complete human genomes from saliva is an even greater priority for our Australian Indigenous genomics research program[23]. Collection of blood during visits to remote Aboriginal communities is impractical and blood is culturally sensitive, thus limiting us to work with saliva[23]. We hope that our demonstration of scalable, affordable genome assembly from saliva – with assembly quality at least on par with the HPRC – may open the door to even greater diversity and inclusion in the growing pangenome field[30]. Applications for Cornetto are not limited to human genomics, as we have shown here by producing highly complete genomes for diverse non-human vertebrates, including three critically endangered, endangered or threatened endemic Australian species. By improving costs, sample input requirements and providing higher quality reference genomes than previously possible, we hope Cornetto will empower other genomically-informed conservation initiatives, similar to the orange-bellied parrot[26], described above. We provide all laboratory and computational methods for Cornetto assembly as a free, open source resource to streamline, improve and democratise genome assembly.

ETHICS AND INCLUSION

Saliva samples were collected from individuals not known to be affected by inherited disease under Human Research Ethics Committee (HREC) approval 95179. Whole blood was collected from patients with FSHD under HREC/2019/ETH12538 and patients with *MUC1*-ADTKD under HREC/18/RPAH/726, HREC/83945/RCHM-2022 and HREC/16/MH/251. Relevant approvals for non-human vertebrates are provided in **Supplementary Note 3**.

METHODS

Cornetto adaptive sequencing and assembly method

Cornetto is a new experimental paradigm in which the genome assembly process is integrated with ONT ReadUntil programmable selective sequencing (also known as 'adaptive sampling'). Cornetto encompasses both laboratory and computational protocols, all of which are described here. Computational steps are executed using the open source Cornetto software package (https://github.com/hasindu2008/cornetto) in addition to other third-party open source software. Most importantly, we have used the excellent software *hifiasm*[7] for generating assemblies and *Readfish*[15] for executing ONT targeted sequencing. *Hifiasm* can be run on the user's preferred machine using commands provided in **Supplementary Note 4** or Cornetto online documentation. *ReadFish* must be executed on the computer that runs ONT sequencing experiments, using reference files generated by Cornetto. Cornetto is also compatible, in principle, with ONT's in-built 'adaptive sampling'

application, which is configured within MinKNOW, and with alternative assembly software, however, these have not been extensively tested.

The Cornetto method starts with a primary assembly of moderate quality, generated with PacBio HiFi or ONT LRS data. For a human genome, LRS data from a single PacBio SMRT cell or a single ONT PromethION flow cell typically yields a suitable base assembly. Our recommended protocols for HMW DNA extraction, DNA shearing, size selection and library preparation, for both PacBio and ONT, are outlined in detail below. A human base assembly is created from the initial LRS data using *hifiasm* (versions and commands in **Supplementary Note 4**).

After generating a base assembly, the available LRS data is realigned to this assembly (using *minimap2*[31]) to assess regional coverage and mappability. Cornetto software is then used to identify assembly regions that are not yet confidently resolved. These are defined as: extended regions of low or high coverage depth (< 40% or > 250% of genome average); low mappability (where coverage depth in uniquely aligned MapQ20+ reads is < 40% of total mean coverage); low assembly quality ('lowQ' regions of > 8 Kbase output by *hifasm*); short primary contigs (< 800Mbase) and; regions adjacent to the end of a primary contig (within 200 kb). For diploid assemblies, Cornetto additionally identifies unphased regions in the assembly. To do so, all contigs from haplotype 1 and haplotype 2 (output by default by *hifiasm*) are aligned to the primary assembly to identify any region that is not spanned by a contig in both haplotypes. These unphased regions are combined with the other labelled regions above; coordinates of all regions are then extended with 40 kb buffers in either direction before merging all overlapping and adjacent features (within 200 kb). All sequences outside of these merged regions, which typically encompass ~10-20% of the genome, are considered to be already confidently assembled and will not benefit from additional LRS data. Hence these regions are considered 'boring bits' and printed to a standard coordinate file 'boringbits.bed' and corresponding file 'boringbits.txt', which is used for ReadFish configuration. Commands to execute this process are provided in **Supplementary Note 4**.

Next, the user should perform ONT sequencing with *ReadFish* (or the ONT adaptive sampling app) configured to reject reads originating from any of the boring bits within the initial assembly. The relevant base assembly is provided as a reference, after indexing with minimap2. Commands, software versions and a template with parameters for ReadFish configuration are provided in **Supplementary Note 4**.

After configuring and launching ReadFish (or the ONT adaptive sampling app within MinKNOW) the user should load their sequencing library onto their ONT flow cell and initiate the sequencing experiment. When starting with a base assembly of PacBio HiFi data, we recommend to run ONT 'duplex' sequencing because HiFi and duplex data are sufficiently similar in accuracy to be co-mingled during assembly. If starting from a base assembly generated with ONT simplex data, the user should continue with ONT simplex data. Protocols used during our study for both ONT sequencing options, including basecalling, are outlined in detail below.

To maximise yields during ONT sequencing, it is standard practice to pause the sequencing process at regular intervals, wash the flow cell with a nuclease solution, reload with fresh library, then resume sequencing. During a Cornetto experiment, the user should take advantage of these pause points in order to update their assembly and regenerate the boringbits reference files. When updating the reference files at each pause point, Cornetto uses the same rules as above, with the exception that poor coverage and mappability rules are not applied beyond the first cycle (because adaptive sequencing introduces uneven coverage which would conflict with these rules). When working with saliva samples, nonhuman contigs may also be added to the list of boringbits for rejection at this point (see below). This process serves to focus ongoing data generation onto an increasingly small and challenging portion of the genome that remains unassembled. We typically perform 3 or 4 Cornetto cycles for a single ONT flow cell, with pause points at ~24 hr, ~48 hr and ~72 hr (if the flow cell remains viable). Each new assembly should be generated by aggregating all existing and new data (see **Supplementary Note 4**). At the end of the process, the user should obtain a final assembly encompassing LRS data from all previous steps. The Cornetto software also provides a simple wrapper script used to evaluate this assembly with a range of standard metrics. The evaluation process run by Cornetto and employed during this study are outlined in

detail below with software versions and commands in **Supplementary Note 4**. Specifics of the process used for FSHD and ADTKD patients are outlined in **Supplementary Note 1** and **2**, respectively. Specifics of the process used for non-human vertebrates are outlined in **Supplementary Note 3**.

High-molecular weight DNA extractions

High–molecular weight (HMW) genomic DNA was extracted from cultured cells from human reference sample hg002, obtained from the Coriell Institute for Medical Research (B-Lymphocyte cell line GM24385), and peripheral blood samples from patients with FSHD and ADTKD, using the PacBio Nanobind CBB kit (102-301-900) or PacBio PanDNA kit (103-260-000) according to the manufacturer's protocol. For saliva experiments, ~5mL of saliva was collected from healthy donors using Oragene self-collection kits (DNA Genotek) and stored at room temperature. Saliva was extracted using the PacBio Nanobind CBB kit with an optimised protocol that is now supported by the manufacturer (103-544-000). For experiments involving non-human vertebrates, HMW DNA was extracted from blood for the petrel, cichlid and turtle and from snap frozen liver tissue for the orange bellied parrot (see **Supplementary Note 3**). For blood samples, approximately 20-70 ul was used as input (or equivalent for ethanol stored blood samples) and these were extracted as per the PacBio Nanobind protocol for nucleated blood. For the liver tissue, 20 mg was used as input and this was extracted using the PacBio Nanobind kit standard Dounce homogenizer protocol. QC checks were performed on all extracted DNA samples using a ThermoFisher NanoDrop (purity), ThermoFisher Qubit (DNA concentration) and Agilent Femto Pulse (Genomic DNA 165 kb Kit; DNA fragment size profiles).

Long-read sequencing methods

For PacBio sequencing experiments, DNA samples were sheared to average fragment lengths of 15–24 Kb using a Diagenode Megaruptor 3 with Shearing Kit at a speed of 30 or 31. Sheared DNA was cleaned, concentrated, then subject to PacBio SMRTbell library preparation, all according to the manufacturer's protocol. Prepped libraries were size selected with a 35% v/v dilution of AMPure PB beads at a ratio of 2.9x (i.e. 50 ul of sample : 145 ul 35% beads) or using a PippinHT (Sage Science) with a 10 kb cut-off, followed by ABC loading procedure and sequencing on a PacBio Revio instrument with 30 hour movie time.

For ONT sequencing experiments, DNA samples were sheared to average fragment lengths of 43–66 Kb using a Diagenode Megaruptor 3 with Shearing Kit and a shearing speed of 27. Sheared samples were treated with PacBio Short-Read Eliminator kit to deplete fragments < 10 kb. ONT libraries were then prepared from ~5–9 µg of sheared HMW genomic DNA using a ligation prep (SQK-LSK114). The resulting libraries were loaded on an ONT PromethION R10.4.1 flow cell (FLO-PRO114M) or a PromethION high-duplex flow cell (FLO-PRO114HD) and sequenced on a PromethION instrument (P2 Solo or P48). Sequencing experiments were typically run for 72 hours, with washes (EXP-WSH004) and library reloading performed at approximately 24 and 48-hour time points. Where flow cells were still viable, an additional wash was performed at 72 hours, followed by a further 24 hr runtime. For Cornetto experiments, live target selection/rejection was executed during the run by the *Readfish*[15] software package with commands and configuration parameters in **Supplementary Note 4**.

Raw ONT sequencing data was converted from POD5 to BLOW5 format[32] in real-time during sequencing. At pause-points during sequencing, or after completion of a run, data was base-called using *slow5-dorado* (v0.8.3; https://github.com/hiruna72/slow5-dorado) with a recent 'super-accuracy' model (dna_r10.4.1_e8.2_400bps_sup@v5.0.0) and a qscore cut off of 10 (--min-qscore 10). For high-duplex flow cells, duplex basecalling was run using *slow5-dorado* (v0.3.4; dna_r10.4.1_e8.2_400bps_sup@v4.2.0) and non-duplex reads were removed prior to downstream analysis. For ONT-only Cornetto experiments, simplex reads were filtered to exclude reads shorter than 30kb (seqkit-v2.3.0)[33], to avoid excessive coverage within on-target regions (noting that this filtering was not performed on reads used to generate the base assembly).

Saliva assemblies and nonhuman reads

For human saliva samples, non-human reads were removed before running *hifiasm* to generate the base assembly. This was performed by running *Centrifuge*[34] on the FASTQ input file (see **Supplementary Note 4**). Additionally, non-human contigs were appended to the reference assembly and boringbits during each cornetto iteration, which facilitates rejection of nonhuman DNA during subsequent sequencing. To do this, non-human contigs must be identified by assembling all the reads in the base assembly (both human and non-human) with *hifiasm*, and then running *Centrifuge* on the assembly. Any contig in this assembly not assigned with the Homo sapiens species code and covered by a minimum of 100 reads are included (see **Supplementary Note 4**). Centrifuge index used was Bacteria, Archaea, Viruses, Human (compressed) index (p_compressed+h+v).

Evaluating genome assemblies

To evaluate human genome assemblies, all contigs are first aligned to the Q100 T2T-hg002 (https://github.com/marbl/HG002) paternal haplotype with chrX added, using *minimap2* 2.24 with preset *asm5* and *--eqx -c* options. Any contigs in the assembly whose sum of aligned lengths for '-' is greater than for '+' with respect to the reference contigs, are reverse complemented. Dot plots are generated using the *minidot* tool in the *miniasm* repository[35]. Telomeres are identified in assembled contigs using the telomere analysis script from the VGP project with default options. To obtain per-chromosome metrics of an assembly, each contig in the the assembly is assigned to the corresponding contig in the hg002-paternal reference based on the contig that was most aligned with. Contig L_{90} values are defined as the number of contigs encompassing >90% of the reference sequence for a given chromosome. A chromosome is considered to be complete if it has a L_{90} = 1 and has a telomere detected at both contig ends. All these operations are carried out using wrapper scripts provided in the Cornetto repository (see **Supplementary Note 4**).

To evaluate assembly contiguity, we generated Nx plots by calculating contig sizes and cumulative assembly size for each additional contig, sorted from largest to smallest. N_{50} and N_{90} values are the minimum contig size for which 50% and 90%, respectively, of the genome is assembled into contigs larger than N Mbase. Compleasm 0.2.6[36] was used for calculating the BUSCO scores with *lineage* set to *primates* for human; *actinopterygii_odb10* for cichlid; and, *tetrapoda_odb10* for both birds and turtles. *Yak* (v0.1) was used to calculate the QV, hamming error and switch error rates for assemblies. For QV value, separate k-mer count indexes are created using *yak count* with k-mer sizes 21 and 31 (*-k* option) using the Q100 T2T-hg002 reference genome which includes both paternal and maternal haplotypes. These indexes are used with *yak qv* subtool to calculate the QV. For calculating hamming and switch errors, first the parental yak indexes for HG002 were downloaded from the human-pangenomics project (https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=submissions/6040D518-FE32-4CEB-B55C-504A05E4D662--HG002_PARENTAL_YAKS/HG002_PARENTS_FULL/). Then *yak trioeval* was used. Example commands are provided in **Supplementary Note 4.**

During evaluation, primary assemblies the user may optionally refine the assembly by retaining complete chromosomes from earlier cornetto cycles, which are sometimes broken during subsequent cycles. The Cornetto software contains a script to execute this, using the following logic. Suppose the base assembly is called asm-0.fasta and the cornetto iterations are named asm-1.fasta, asm-2.fasta, asm-3.fasta, ..., asm-n.fasta. Starting from asm-1.fasta, asm-2.fasta, asm-3.fasta, ..., asm-n.fasta are iterated until any contigs are found longer than the expected minimum chromosome size and have telomeres in both ends (considered to be 'complete chromosomes'). Suppose complete chromosomes are found in asm-k.fasta. Now such contigs are extracted from asm-k.fasta into a file called asm.fasta. Now starting from asm-(k+1).fasta, assemblies are iterated till asm-n.fasta (including asm-n.fasta). At each iteration, any complete chromosomes in the assembly are mapped to asm-k.fasta. Any newly found complete chromosomes are appended to asm.fasta (those contigs which map <50% of their length to a contig into asm.fasta). At the last iteration (asm-n.fasta), any other contigs

(that are not considered complete chromosomes) are also mapped to asm.fasta. Any contigs which map <50% of their length to a contig are appended into asm.fasta. At the end, asm.fasta is the final curated primary assembly.

DATA AVAILABILITY

Raw sequencing data is deposited at ENA project PRJEB86853 and will be made public at the time of publication. Raw datasets for human participants may be accessed upon reasonable request. Genome assemblies are available at Dryad (10.5061/dryad.kkwh70sfr) and will be made public at the time of publication. Genome assemblies for human participants may be accessed upon reasonable request. The following publicly accessible datasets were also used in this study:

hg002 PacBio HiFi data from the T2T Consortium:

https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/scratch/HG002/sequencing/hifirevio/m84005_220827_014912_s1.hifi_reads.fastq.gz

hg002 ONT duplex data from the T2T Consortium:
https://human-pangenomics.s3.amazonaws.com/index.html?prefix=submissions/0CB931D5-AE0C-4187-8BD8-B3A9C9BFDADE--UCSC_HG002_R1041_Duplex_Dorado/Dorado_v0.1.1/stereo_duplex/

T2TC-ONT-only assembly (Koren et al 50x duplex + 30x UL + PoreC) assembly:

https://obj.umiacs.umd.edu/marbl_publications/duplex/HG002/asms/duplex_50x_30xUL_poreC.tar.gz

CODE AVAILABILITY

Cornetto software is open source and freely available: https://github.com/hasindu2008/cornetto

All original code has been deposited at Zenodo and is publicly available: 10.5281/zenodo.15075988.

ACKNOWLEDGEMENTS

The authors acknowledge the traditional custodians of the land upon which the orange-bellied parrot, western saw-shelled turtle, Gould's petrel and redstriped eartheater cichlid reside, as well as the custodians of the historic range of each species. We thank Deborah Bower and Yuna Kim for involvement in specimen collection for the turtle and petrel, respectively. We thank Priam Psittaculture Centre for sampling of the parrot. This project was undertaken with services from the National Computational Infrastructure (NCI). We thank Tim Ho for expert technical support and allowing us to use Garvan Institute data science infrastructure sometimes in quite exotic ways.

We acknowledge the following funding support: Australian Medical Research Futures Fund grants, 2023126, 2041648, 2025138, 2008249, National Health and Medical Research Council (NHMRC) grant 2035037 (to I.W.D.), Australian Research Council (ARC) DECRA Fellowship DE230100178 and ARC Discovery Project DP230100651 (to H.G.). A.J.M was supported by a Queensland Health Advancing Clinical Research Fellowship. L.S. is supported by the ARC Centre of Excellence in Innovations in Peptide and Protein Science (CE200100012). Work on the orange-bellied parrot was supported by Australian Biocommons which is enabled by National Collaborative Research Infrastructure Scheme via Bioplatforms Australia Threatened Species Initiative funding; DNRF143. The views expressed herein are those of the authors and are not necessarily those of the Australian Government or the ARC, NHMRC or MRFF.

AUTHOR CONTRIBUTIONS

H.G., H.R.P. & I.W.D. conceived the project.

H.G., K.J., & I.W.D. developed the Cornetto software.

I.S., J.M.H., M.R., T.R. & Y.H. performed laboratory experiments.

I.S., D.Y., Y.H., A.J.M., E.S., K.R.K. & A.C.M. recruited, collected and processed patient samples.

J.M.H., L.R., L.W.S., C.J.H., L.S., O.B., R.C.R.N., L.A.S.N., A.L.C. & A.G. coordinated, collected and processed non-human samples.

H.G., I.S., J.M.H., A.L.M.R., L.W.S., D.Y., H.C., H.R.P. & I.W.D. performed bioinformatics analysis.

H.G., A.L.M.R., D.Y., A.C.M. & I.W.D. prepared the figures and tables.

H.G., I.S. & I.W.D. wrote the manuscript, with contributions from all co-authors.

DECLARATIONS

I.W.D. manages a fee-for-service sequencing facility at the Garvan Institute and is a customer of Oxford Nanopore Technologies and Pacific BioSciences but has no further financial relationship. H.G., A.L.M. and I.W.D. have received travel and accommodation expenses from Oxford Nanopore Technologies. The authors declare no other competing financial or nonfinancial interests.

REFERENCES

1. Li, H. & Durbin, R. Genome assembly in the telomere-to-telomere era. *Nature Reviews Genetics* **25**, 658–670 (2024).
2. Completing human genomes. *Nat Methods* **19**, 629 (2022).
3. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
4. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
5. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
6. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).
7. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
8. Koren, S. *et al.* Gapless assembly of complete human and plant chromosomes using only nanopore sequencing. *Genome Res* **34**, 1919–1930 (2024).
9. Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nature Biotechnology* **41**, 1474–1482 (2023).
10. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
11. Mc Cartney, A. M. *et al.* Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nature Methods* **19**, 687–695 (2022).
12. Wagner, J. *et al.* Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol* **40**, 672–680 (2022).
13. Kronenberg, Z. N. *et al.* Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nat Commun* **12**, 1935 (2021).
14. Dahn, H. A. *et al.* Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing. *Gigascience* **11**, (2022).
15. Payne, A. *et al.* Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat Biotechnol* **39**, 442–450 (2021).
16. Stevanovski, I. *et al.* Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Sci Adv* **8**, eabm5386 (2022).
17. Miller, D. E. *et al.* Targeted long-read sequencing identifies missing disease-causing variation. *Am J Hum Genet* **108**, 1436–1449 (2021).
18. Rudaks, L. I. *et al.* Targeted long-read sequencing as a single assay improves the diagnosis of spastic-ataxia disorders. *Ann Clin Transl Neurol.* (2025) doi: 10.1002/acn3.70008.
19. Preston, M. K., Tawil, R. & Wang, L. H. Facioscapulohumeral Muscular Dystrophy. in *GeneReviews* (University of Washington, Seattle, 2020).
20. Autosomal dominant tubulointerstitial kidney disease. *Nat Rev Dis Primers* **5**, 61 (2019).
21. Kirby, A. *et al.* Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. *Nat Genet* **45**, 299–303 (2013).
22. Website.
23. Reis, A. L. M. *et al.* The landscape of genomic structural variation in Indigenous Australians. *Nature* **624**, 602–610 (2023).
24. Phillips, A. M. Only a click away - DTC genetics for ancestry, health, love…and more: A view of the business and regulatory landscape. *Appl Transl Genom* **8**, 16–22 (2016).
25. Abraham, J. E. *et al.* Saliva samples are a viable alternative to blood samples as a source of DNA for high throughput genotyping. *BMC Med Genomics* **5**, 19 (2012).
26. Silver, L. W. *et al.* Temporal Loss of Genome-Wide and Immunogenetic Diversity in a Near-Extinct Parrot. *Mol Ecol* e17746 (2025).
27. Stanojević, D., Lin, D., Nurk, S., de Sessions, P. F. & Šikić, M. Telomere-to-Telomere Phased Genome Assembly Using HERRO-Corrected Simplex Nanopore Reads. *bioRxiv* 2024.05.18.594796 (2024) doi:10.1101/2024.05.18.594796.
28. Bzikadze, A. V. & Pevzner, P. A. Automated assembly of centromeres from ultra-long error-prone reads. *Nat Biotechnol* **38**, 1309–1316 (2020).
29. Bleyer, A. J., Živná, M., Kidd, K. & Kmoch, S. Autosomal Dominant Tubulointerstitial Kidney Disease – MUC1. in *GeneReviews* (University of Washington, Seattle, 2021).
30. Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nat Rev Genet* **21**, 243–254 (2020).

31. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
32. Gamaarachchi, H. *et al.* Fast nanopore sequencing data analysis with SLOW5. *Nat Biotechnol* **40**, 1026–1029 (2022).
33. Shen, W., Sipos, B. & Zhao, L. SeqKit2: A Swiss army knife for sequence and alignment processing. *iMeta* **3**, e191 (2024).
34. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* **26**, 1721–1729 (2016).
35. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
36. Huang, N. & Li, H. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics* **39**, (2023).