C e nding a h : Chen-G ang Li , Zh Wang, S a e Ke Lab a f Mic bial Me ab li m, Bi -X In i e , Ke Lab a f he Gene ic f De el men al and Ne chia ic Di de f he Mini f Ed ca i n, J in In e na i nal Re ea ch Lab a f Me ab lic & De el men al Science f he Mini f Ed ca i n, Sch l f Life Science and Bi echn l g , Shanghai Jia T ng Uni e i , Shanghai 200240, China. Tel: 86-21-34205125; Fa : 86-21-34208028; E-mail: cg.li @ j .ed .cn (C.-G. Li); Tel: 86-21-62933338; Fa : 86-21-62932059; E-mail: h ang@ j .ed .cn (Z. Wang)

Background: *de novo*
H , . C
. P B H F . O N (ON)
P ION - - , -

Methods: 455 (7 . . 4 13. .) 88 (4
11.) *de novo* . . 288C - ON
H F . . , . . Q A (Q A), B -
C . O . (B CO) C (C .). I , .

Results: . . , - . . F ON . . ,
F . . C . P P M
. , . , . F H F . , F .
N D . , . . E -
. . ON (>80) H F (>20) . .

– . . , - , , , , . .

The high- h gh and l ng- ead f ne -gene a i n e encing echn l gie enabled he e encing f en i e gen me a an n eceden ed eed, hich ha e l i ni ed bi l g in he a decade n nl f lab a e ea ch b f e le' dail life [1]. The hi d-gene a i n e encing (TGS), Pacific Bi cience (PacBi) and O f d Nan e Techn l gie (ONT) a e eal- ime, l ng- ead gene a ed, ingle- m lec le e encing la f m , hich can e c me he h c ming f ec nd-gene a i n e encing (SGS) echn l g ch a ela i el h ead , e ence- de enden bia e , inf ma i n l [1], ha g ea l im ed he c n ini f *de novo* gen me a embl [2].

La el , high- h gh high-fideli (HiFi) ead , b ained b PacBi e el II em i h he ci c la c n en e encing m de, ned l ng ead (>10 kb) and high e -ba e acc ac (>99.9%). Unlike SGS and PacBi ' ical m ni ing em ha el n DNA l me a e ead ba e e ence , ONT e encing iden ifie DNA ba e b mea ing he change in elec i- cal c nd c i i gene a ed a DNA and a h gh a bi l gical e ha make i gene a e l al ng ead . The m c ing-edge e e en a i n i P me hION la f m ha can d ce 7 Tb ead e n i h an a e age e encing eed f ~430 ba e / and N50 > 20 kb [3], hich make i ible achie e da a ickl and e c me he c n in f e e i i e egi n

Xue Zhang i a Ph.D. candida e in he Sch l f Life Science and Bi echn l g a Shanghai Jia T ng Uni e i . He e ea ch in e e a e in c m a a i e gen mic anal i and in eg a i e mic anal i .
Chen-Guang Liu i an A cia e P fe in he Sch l f Life Science and Bi echn l g a Shanghai Jia T ng Uni e i . Hi e ea ch in e e a e in bi efine f m lign cell l e and me ab lic enginee ing.
Shi-Hui Yang i a P fe in he Sch l f Life Science a H bei Uni e i . Hi e ea ch in e e a e in me ab lic enginee ing, n he ic bi l g , and ene able bi d c .
Xia Wang i an A i an P fe in he Sch l f Life Science a H bei Uni e i . He e ea ch in e e a e in me ab lic enginee ing, n he ic bi l g , and ene able bi d c .
Feng-Wu Bai i a P fe in he Sch l f Life Science and Bi echn l g a Shanghai Jia T ng Uni e i . Hi e ea ch in e e a e in bi efine , bi ce enginee ing, and me ab lic enginee ing.
Zhuo Wang i an A cia e P fe in Bi -X In i e , Sch l f Life Science and Bi echn l g , Shanghai Jia T ng Uni e i . He e ea ch in e e a e in bi inf ma ic , ne k m deling and in eg a i e mic anal i .
Received: N 27, 2021. **Revised:** M 26, 2022. **Accepted:** M 31, 2022
A ( ) 2022. P O P . A . F P , : . . @ .

c n c a c n in high- ali gen me e ence, e eciall in he a embl f c m le gen me ch a high he e g i , high e e i i n and la ge gen me .

Al h gh l ng ead d beneficial gen me c n- c i n, he ingle ba e e a e f ONT and HiFi e encing i ill highe han SGS. The e a e mainl a egie f ed cing e in ONT gen me a emblie ; he fi i c ec he and m e f l ng ead bef e gen me a embl h gh high c e age, and he ec nd i li h he d af e ence af e a embl hich e called ' li hing' e .

Recen l , a a ie f ne gen me a embl and l- i hing me h d ha e eme ged. Can i a cce f Cele a a emble and in d ce he ada i e e la- ing a eg ba ed n *tf-idf* eigh ed MinHa h and he a e a embl g a h c n c i n ha a id c lla- ing di e ged e ea and ha l e [4]. I a la el m difie d f HiFi ead and named HiCan [5], hich ha h m l me c m e i n, e la -ba ed e c ec- i n and agg e i e fal e e la fil e ing e . NECAT i an e c ec i n and *de novo* a embl l f ONT l ng n i ead [6]. Simila Can , i fi c ec he a ead and hen e abli he he a embl , b Neca e an ada i e ead elec i n and - e g e i e me h d ickl c ec ONT ead high acc ac . Ne Den [7] i a ing g a h-ba ed *de novo* a emble f m li i e l ng ead ch a C n in L ng-Read (CLR), HiFi and ONT. I e he 'c ec - hen-a emble' a eg f ONT ead b n c ec i n e f HiFi ead . Fl e [8] i al c m a ible i h ONT and HiFi ead and fi gene a e he e - ne di j in ing , hen c nca ena e all di j in ing in a ingle ing in an a bi a de and c n c he a embl g a h and finall e l e he g a h b ain he acc a e c n ig . Hifia m [9] i a ecen l bli hed a emble ha ake ad an age f HiFi ead fai hf ll e l e he ha l e inf ma i n in a ha ed a embl g a h. Minia m c n c he ing g a h b e la ing a e f ead . I nl e f m he la e f he e la la - c n en alg i hm, hich i diffe en f m Can Neca . Unlike he ab e l ng- ead a em- ble , Unic cle [10] can e h - ead- nl , l ng- ead- nl h b id ead f a embl . In he h b id a embl ce , i b ild an ini ial a embl g a h f m h ead ing SPAde and hen im lifie he g a h ing inf ma i n f m h and l ng ead . Pil n [11] and Rac n a e li he ha e h ead and l ng ead f c ec i n, e ec i el . Rac n i ec mmended c ec Minia m' d af a embl [12], e al de ign he ONT ead a embl i eline named MiniRac n ha c mbine he Minia m and hi d nd Rac n i e a i n. Medaka i he fi ne al ne k-ba ed li hing l de el ed b ONT. Ne alP li h [13] i he la e li he ba ed n alignmen ma i c n c i n and h g nal Bi-di ec i nal Ga ed Rec en Uni (Bi-GRU) ne k .

P e i k n he c m a i n f gen me a embl l f ONT ead m l ed ka e , ch a *Escherichia coli* [14], i [15] and a h genic bac e ia [16, 17], and m l ed im la ed da a e c n c l - ali e ence . Gi dan *et al.* [18] ed ea a e encing ma e ial f *de novo* a embl c m a i n in 2017, b he e encing de h f he ONT da a e a e l , nl ~30X and he l ed e e - da ed, hich c ld n g ide he ch ice f ad anced l f he gen me c n c i n. Recen l , he c m a i n f a embl n HiFi ead a ca ied f *E. coli*, *Drosophila ananassae* [19] and ice [20]. H e e , HiFi ha n been c nd c ed and e al a ed in ea , n men- i n he c m a i n f all a ailable a emble . The e- f e, elec he ad anced gen me-b ild i eline and a ia e e encing de h f e ka e emain gen l be in e iga ed.

In hi d , e c m ehen i el e al a ed he infl ence f e encing me h d , a embl l , l- i hing l and e encing de h n e ka ic gen me c n c i n i h high-c e age ONT, HiFi and BGISEQ ai ed-end da a e , ing m del gani m ea a he e e en a i e. Thi i n nl he fi ime ac he ide c e age a 800X n he ONT da a e b al he fi HiFi da a e elea e f ea . F he ONT da a e , he c mbina i n f 7 a embl me h d (Can , Fl e, Neca , Minia m, MiniRac n, Ne Den and Unic cle) and 4 li hing me h d (Medaka, Pil n, Ne alP li h and Medaka_Pil n) a ca ied n 13 diffe en de h b e . While f he HiFi da a e , 4 la e l , hifia m, Ne Den , Fl e, and HiCan , e e a lied n 11 diffe en de h b e i h i h Pil n li h. The gen me ali and c m ing e f mance f each i eline a e al a ed de e mine he a ia e kfl and e encing de h. In addi i n, e al e enced and a embled diffe en gene a f ea ain , *Saccharomyces cerevisiae* CICC-1445 (SC) and *Schizosaccharomyces pombe* FLO-DUT (SP), e ie e he high- ali gen me b ing he anding kfl . The ad anced benchma k e f da ed l ba ed n high-c e age ONT and HiFi da a e ide al able g idance f TGS gen me c n c i n f m ld, mic algae and e en c m le gen me , ch a h man, be e ce i nall e en l in nde anding he f ndamen al mechani m f e ka ic.

Yeast strain and growth conditions

Saccharomyces cerevisiae ain S288C (ATCC 204508) and *S. cerevisiae* CICC-1445 e e cha ed f m he Ame- ican T e C l e C llec i n (ATCC) and China Cen e f Ind ial C l e C llec i n (CICC), e ec i el ; *S. pombe* FLO-DUT a e e ed b lab a . The ea ain a ec l ed n Yea e ac Pe ne De e medi m (YPD) la e (10 g/l ea e ac , 20 g/l e ne, 20 g/l gl c e and 20 g/l aga) a 30 °C f 24 h. A ingle c l n a ac i a ed and hen c l ed in YPD media a 30 °C and 150 m and am led hen cell e e g ing in he l g ha e i h O ical Den i $(OD)_{600}$= 0.8 ~ 0.9.

Genome sequencing and subsets generation from raw full datasets

Gen mic DNA f h ee ain a e ac ed, e ed and e enced gene a e ONT (P me hION) and BGISEQ ead and HiFi ead a BGI (Shen hen, China) and Pe - nalbi (Shanghai, China), e ec i el . *Saccharomyces cerevisiae* CICC-1445 and *S. pombe* FLO-DUT ha e n been e enced bef e. He e, e gene a ed he fi c m le e gen me e ence f he e ain . All a da a and h ee a embled gen me h gh imal i eline ha e been l aded in he Na i nal Cen e f Bi echn l g Inf ma i n (NCBI) i h Bi P jec acce i n PRJNA792930, PRJNA792931 and PRJNA792932 f *S. cerevisiae* S288C, *S. cerevisiae* CICC-1445 and *S. pombe* FLO-DUT, e ec i el .

De ailed a i ic inf ma i n ab each e enced da a e i mma i ed in S lemen a Table S1 S3 (ee S lemen a Da a a ailable nline a h ://academic. .c m/bib). S288C ha he dee e e enc- ing de h, ab 800X ONT ead i h N50 f ∼27 kb, 380X HiFi ead i h N50 f ∼21 kb and a e age a e i h 10.29 ime and 240X de h f 2 × 150 b BGISEQ ai ed ead . The a ONT and HiFi ead f S288C e e ma ed he la e g ld- anda d efe ence S288C gen me (GCA_000146045.2) b bmm2 (1.3.0) [21] calc la e he mean-ma ed c nc dance. Al h gh e encing de h a e diffe en f h ee ain , hei ONT (S lemen a Fig e S1A and B, ee S lemen- a Da a a ailable nline a h ://academic. . c m/bib) and HiFi (S lemen a Fig e S1C and D, ee S lemen a Da a a ailable nline a h :// academic. .c m/bib) da a e ha e imila ead leng h di ib i n. T e l e he de endence f he a embl i eline n e encing de h and he effec f ha n he a embl ali , e and ml am led 13 b e i h g adien de h f 10X, 20X, 40X, 60X, 80X, 100X, 120X, 140X, 160X, 320X, 480X, 640X and 800X f S288C ONT da a, 11 b e i h g adien de h f 10X, 20X, 40X, 60X, 80X, 100X, 120X, 140X, 160X, 320X and 380X f S288C HiFi da a and 6 b e i h de h f 10X, 20X, 40X, 80X, 160X and 240X f BGISEQ da a b Se k 1.2. Each ONT HiFi b e ha a imila ead leng h di ib i n (S lemen a Fig e S1E and G, ee S lemen a Da a a ailable nline a h :// academic. .c m/bib) and c inciden ead leng h den i (S lemen a Fig e S1F and H, ee S le- men a Da a a ailable nline a h ://academic. . c m/bib).

De novo assembly and polishing pipelines

De novo a embl and li hing i eline f benchma k e a e h n in Fig e 1. F 11 HiFi b e f S288C, e e f med 4 a emble , HiCan (2.2), Fl e (2.8.3), Ne Den (2.5.0) and hifia m (0.16.1) b defa l e ing . F 13 ONT b e f S288C, e ed 7 diffe en l f l ng- ead a embl : Can (1.9), Fl e (2.8.3), Neca (20200119), Minia m (0.3), Minia m (0.3)/Rac n (1.4.13), Ne Den (2.5.0) and h b id- ead a emble Unic cle (0.4.8). Defa l a ame e e e ed e ce Fl e a e -high ead de h, ' a m-c e age = 50' a e . The n mal m de a ed f he h b id a embl b Unic cle . Then, each a embl a f he li hed b Ne alP li h, Medaka (1.0.1) Pil n (1.23), e ec i el , i h defa l a ame e . Onl ONT ead e e ed in Ne alP li h and Medaka, hile nl BGISEQ ead e e in in Pil n. The m del f ' 941_ m_high_g303' a ed in Medaka. We al c mbined Medaka and Pil n i eline (Medaka_Pil n) b ain high- ali gen me a embl . When l e han 160X, he de h f h ead a c n i en i h ha f l ng ead d ing he a embl ce . In he ca e , all h ead (240X) e e ed.

Genome assembly assessment

We ed QUAST (5.0.2) e al a e he ali f he a emblie gene a ed b diffe en a emble and l- i hing l [22]. All d af a emblie e e c m a ed he la e efe ence S288C gen me. F m e e al me ic , e elec ed he n mbe f c n ig , N50, Mi ma che and Indel e 100 kb i ali e in he main e . We al ed BUSCO [23] a e he gen me ann a i n c m le ene f a emblie . In de cla if he a e f diffe en e e f Ne alP li h, e ed P m i 0.3.9 e he in e i n and dele i n e a e e a a el . F he m e, e ed a ne C m e- hen i e_c e (C_c e) c m i i el e al a e he ali f gen me a emblie f m diffe en i eline in de gi e me ad ice f needed e , hich e e de c ibed in de ail in ec i n 3.2. And e al a ach he a e al a ing al e f m QUAST in he lemen al file f ad anced e ' inf ma i n.

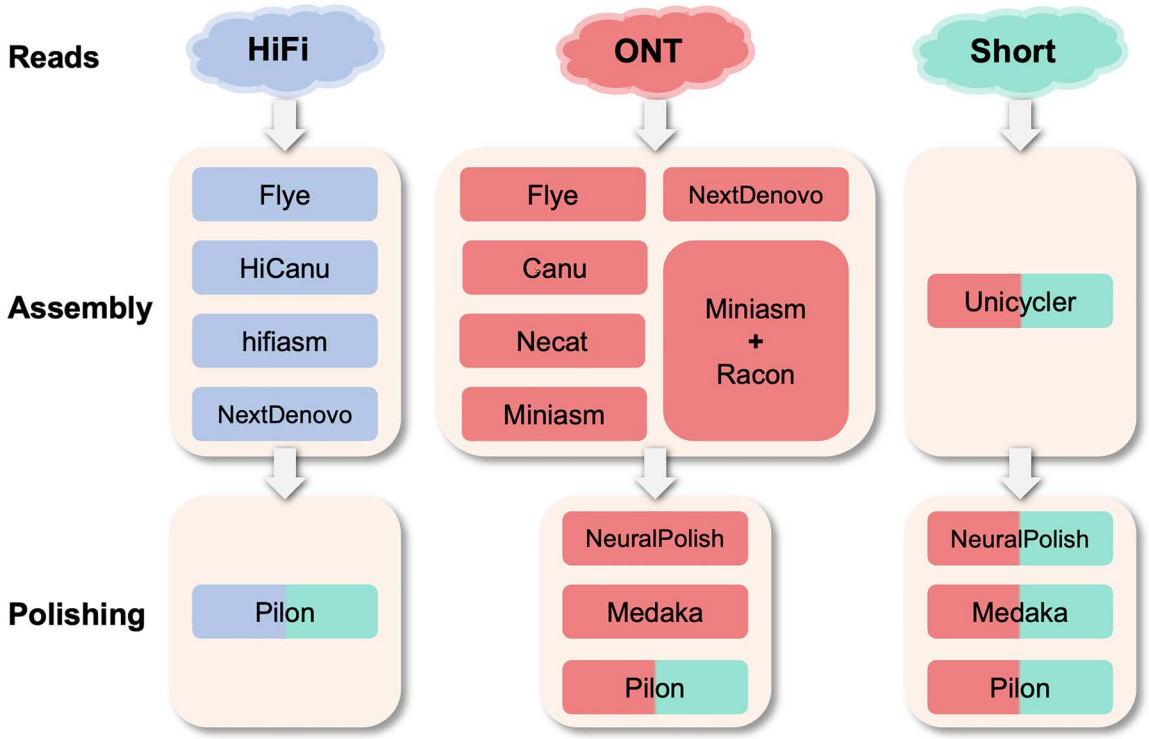
We de igned he im emen a e (IR) f each me ic e al a e he li hing i eline h gh he E a i n (1), he e i mean he de f he da a de h and *n* mean he n mbe f b e hich i 13 f ONT da a. The c efficien i +1 −1 f me ic mean ha he highe he be e (N50) he malle he be e (Indel , Mi ma che , c n ig), e ec i el

$$\mathrm{I} = (\pm 1) \times \frac{1}{n} \sum_{i=1}^{n} \frac{M_{i\_A} - M_{i\_B}}{M_{i\_B}}. \quad (1)$$

Annotation of the assembled genome

De novo gen me c e ann a i n a ca ied b A g (3.3.3) [24]. Then, ann a ed ein e ence e e aligned he S288C efe ence da aba e b BLASP i h a al e = 1e−5 and ma _a ge _e = 1. We calc la ed he e cen age f ann a ed gene b E a i n (2). The al gene n mbe in efe ence a 6002. Venn diag am a d a n b VENNY (2.1) [25] c m a e he gene ann a i n abili f imal i eline .

$$P = \frac{A}{\quad}. \quad (2)$$



. . *De novo* a embl and li hing i eline f benchma king. The da a e ed in each i eline e e e en ed b c l : Blue, HiFi; Red, ONT; G een, h ead f m BGISEQ.

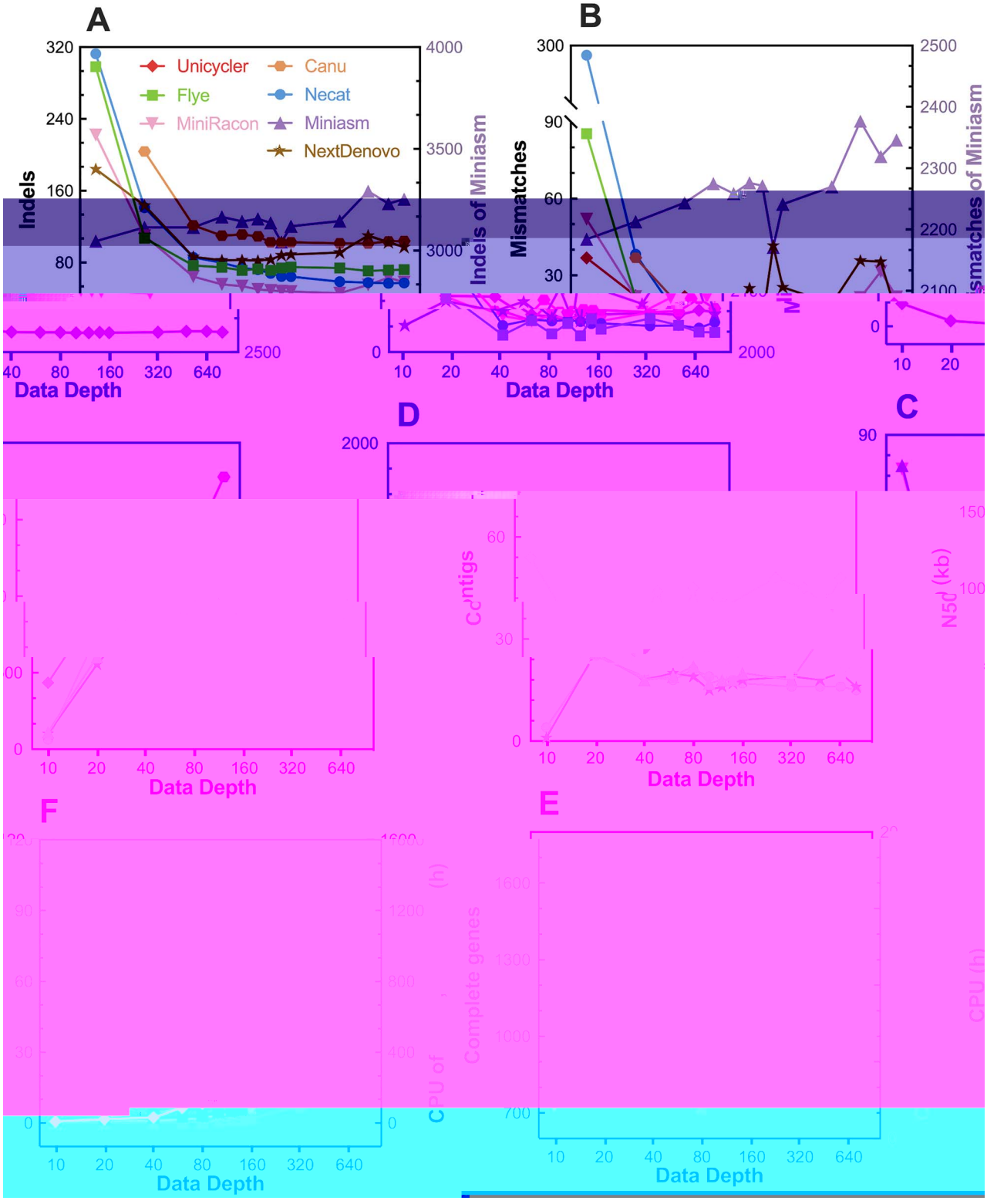
The e f g am f a embl , li hing, e al- a i n and ann a i n e e ided in lemen al file.

De novo assemblers showed significant differences in assembly quality on ONT datasets

We ed he efe ence ain f *S. cerevisiae*, S288C, c m a e he ca abili ie f e en ad anced a embl i eline , Can , Fl e, Neca , Minia m, MiniRac n, Ne Den and Unic cle . The e a embl i eline h ed ignifican diffe ence in a embl ali . Gen me acc ac e eciall he Indel e 100 kb a highl a iable (Fig e 2A). Unic cle ha he m anding e f mance n Indel e 100 kb, i h nl a fe e , f ll ed b MiniRac n, Neca and Fl e. In addi i n Minia m, Can ha he m Indel . M f he a emble ha e ela i el l Mi ma che , hile Fl e and Neca e en ed he l e mi ma ch n mbe and ha e l a iabili ie in diffe en ead de h (Fig e 2B). A f c n ini i , Neca and Ne Den - d ced d af a emblie ha had c n ig n mbe cl e he ch m me n mbe f he efe ence gen me (16 ch m me and 2 la mid) and Fl e ha he b N50 (Fig e 2C and D). Acc ding he e l f BUSCO (Fig e 2E and S lemen a Fig e S2, ee S lemen- a Da a a ailable nline a h ://academic. .c m/bib), Unic cle ha he m c m le e gene n mbe , f ll ed b MiniRac n, Fl e and Neca .

Can ha he lea b c n ig n mbe , i h fe e c n ig hen da a de h a bel 480X and d ama icall inc ea ed hen da a de h a ab e 480X (Fig e 2). Unic cle e f med eake in e m f c n ig i b e i in e m f acc ac . Minia m a embl e en he n b h acc ac and c m le ene ince i d e n' incl de he ba e e c ec i n and c n en e . B BUSCO e al a i n, i nl ha 2 11 c m le e gene , i h a i n f 0.1 0.5% (S lemen a Fig e S2, ee S lemen a Da a a ailable nline a h ://academic. .c m/bib), hich indica e he l f c ec i n and c n en d ing a embl ha a g ea infl ence n he b e en e ence cha ac e i ic anal i .

A he da a de h f he b e inc ea e f m 10X 800X, he c m ing e ce e i ed b all a em- ble inc ea e (Fig e 2F). Minia m and Ne Den c n med he lea CPU ime, f ll ed b MiniRa- c n and Neca . Unic cle c n med m c m a- i nal e ce and k 1536 CPU h i h 800X b e , f ll ed b Can , hile Ne Den i a age-c n med a emble and end m mem (S lemen a Fig e S3, ee S lemen a Da a a ailable nline a h ://academic. .c m/bib).



- . , Main me ic f a emblie n 13 ONT b e i h diffe en da a de h . N mbe f Indel (**A**) and Mi ma che (**B**) e 100 kb, c n ig n mbe (**C**), leng h f N50 (**D**), c m le e gene ' n mbe f m BUSCO (**E**) and c m a i nal ime f diffe en a emble (**F**).

C_score evaluation of assemblers

He e, e in eg a e 5 me ic c m ehen i el e al a e he gen me ali ie f a emblie f m diffe en a emble , incl ding c n ig n mbe (c n ig), he n mbe f mi ma che (Mi ma che), Indel e 100 kb (Indel), N50 leng h in kb and he n mbe f c m le e gene e al a ed b BUSCO (c m le ene). F each me ic, he al e a caled [0, 1] b Min-Ma N mali a i n ac dife en i eline and named Scaled_Me ic (SM) h gh E a i n (3), he e M i he mean f all a ailable b e

Table 1. The mean f me ic al e (M) and c m ehen i e c e (C_ c e) f diffe en a emble n ONT da a e .

	Contigs	N50 (kb)	Mismatches	Indels	Completeness	C_score
Fl e	25.0	942.4	16.3	93.1	1506.7	0.904
Neca	17.2	825.4	33.3	91.8	1419.2	0.824
Ne Den	17.2	798.2	22.3	101.8	1446.2	0.795
MiniRac n	26.5	788.1	19.7	70.9	1512.5	0.713
Can	35.8	834.9	18.4	114.4	1458.1	0.678
Unic cle	40.3	836.6	17.6	4.4	1697.6	0.677
Minia m	26.9	770.8	2265.1	3150.6	5.9	0.116

*E . M . . . , 13 ON . . . (M) , 10 . , 800 . .

me ic (M_i). The M_{min} M_{ma} mean he minim m ma im m M f i eline ha a e be c m a ed.

$$M = \frac{M - M}{M - M}. \quad (3)$$

F high- ali a embl , he me ic N50 and he n mbe f c m le e gene h ld be a high a ible, and he me ic c n ig , Mi ma che , Indel h ld be a l a ible. The ef e, e define a a _C_ c e in E a i n (4) b mming fi e SM , h e c efficien e e e a 1 f he fi me ic −1 f he la e h ee me ic becau e f he i i e and nega i e c n- ib i n f me ic . Then, e e cale he a _C_ c e [0, 1] al b Min-Ma N mali a i n and b ained he C_ c e E a i n (5), he e he _C_ , _ , and _C_ , _ mean he minim m and ma im m he e ical al e f he a _C_ c e, hich i −3 and 2, e ec i el

$$\_C\_\ ,\ = M_{N50} + M_{C}\ ,\ .. - M_{C}\ , - M_{M}. \ . - M_{I}\ . \quad (4)$$

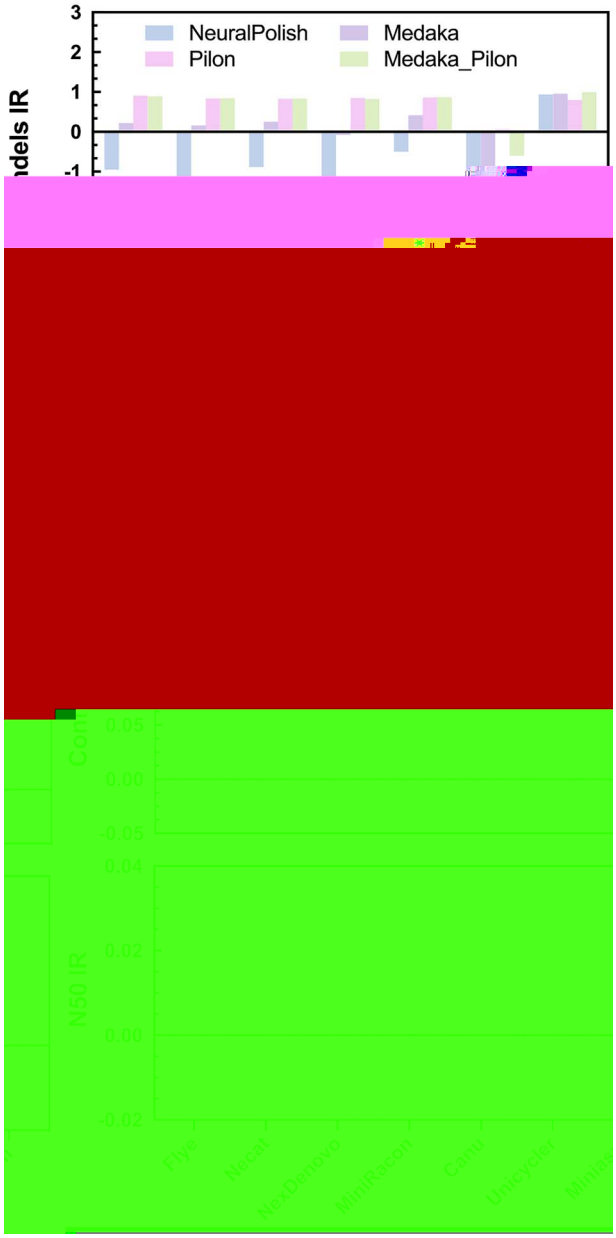
$$C\_\ ,\ = \frac{\_C\_\ ,\ - \left(\ \_C\_\ ,\ \right)}{\_C\_\ ,\ - \ \_C\_\ ,\ }. \quad (5)$$

The me ic mean f he 13 b e i ed calc la e he C_ c e f 7 a emble (Table 1). Fl e ha he be c m ehen i e e f mance, cl el f ll ed b Neca . Be ide , Unic cle ha e cellen acc ac de i e he f agmen ed a emblie and l ng CPU h .

Influence of polishing process on assembly quality

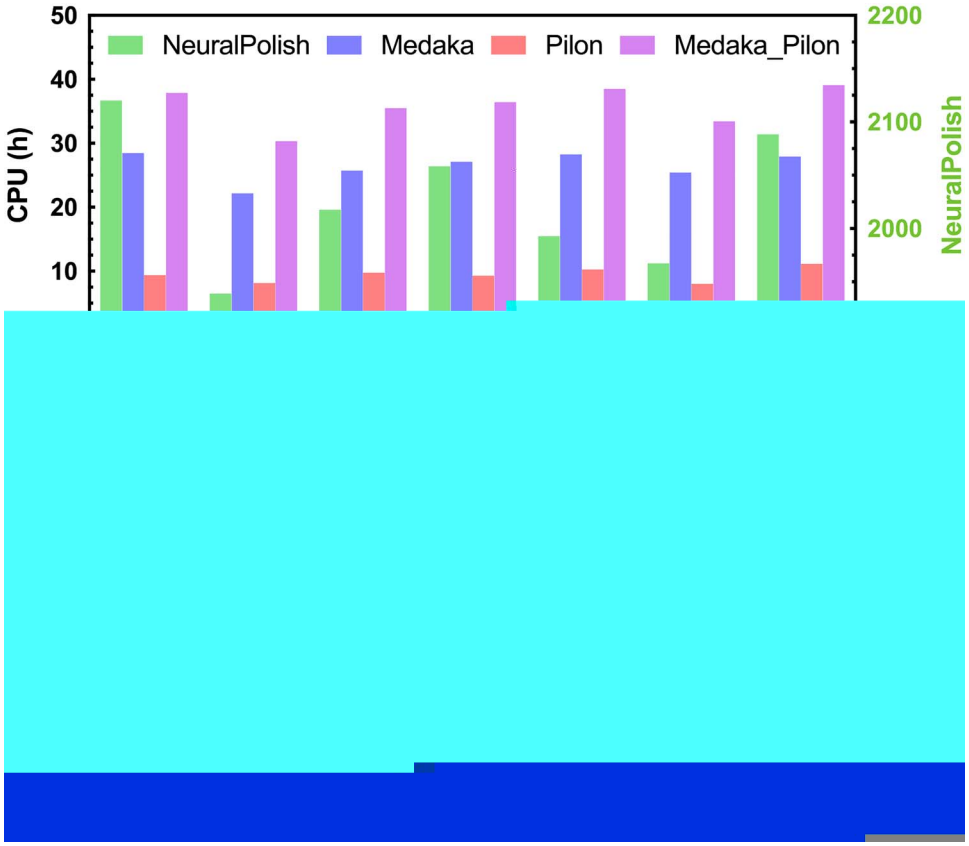
F each b ained d af a embl f ONT b e , e b ained a li hed a embl b ing each n f f li hing i eline (Ne alP li h, Medaka, Pil n, Medaka_Pil n) and hen a e ed i e f mance h gh QUAST (S lemen a Fig e S4, ee S lemen a Da a a ailable nline a h ://academic. . c m/bib) and calc la ed he main me ic IR f each li hing ce (Fig e 3).

Medaka can ed ce he c n ig n mbe a ce ain deg ee and Ne alP li h can im e he N50 me ic ince he e he l ng ead im e he c n ini i . Pil n ha n b i effec n he im emen f c n ini becau e i e he h ead f m SGS f



- . . Me ic IR f a emblie af e he li hing ce .

fine li hing f ba e . H e e, in e m f acc ac , Pil n i he m b li he ed ce Mi ma che and Indel am ng h ee ingle li he , cl el f ll ed b Medaka. Medaka can ed ce he Mi ma che f d af e ence , e eciall f Unic cle . Af e Unic cle ' a emb l and Medaka' li hing, e b ained he

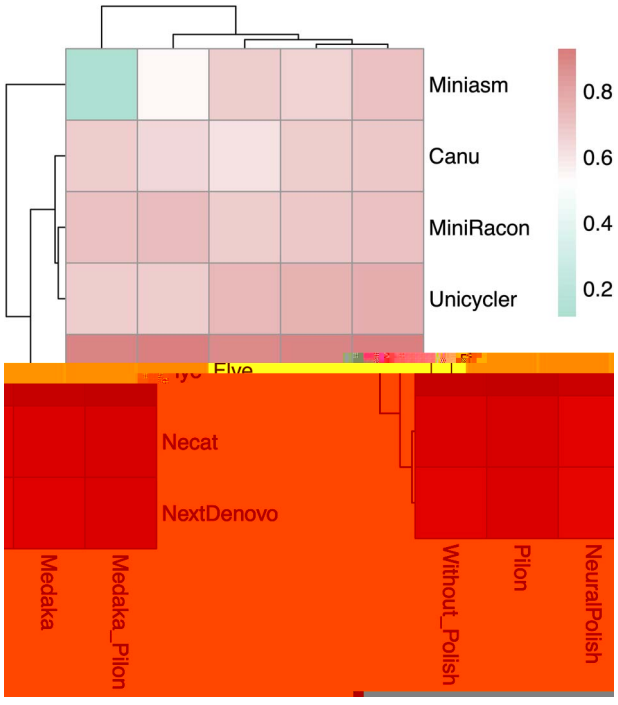


- . C m a inal e ce e ied f li he . The al ei he m f 13 b e i h da a de h .

lea mi ma ched e ence (S lemen a Fig e S4, ee S lemen a Da a a ailable nline a h ://academic. .c m/bib). Ne alP li h h ed a iable e f mance n acc ac . On ne hand, i can ed ce Mi ma che and Indel f a emblie f m Minia m, n he he hand, i eall inc ea e ha f he a emblie , hich i d e he in d c i n f in e i n e a he han dele i n e (S lemen a Fig e S5, ee S lemen a Da a a ailable nline a h ://academic. .c m/bib). I i ec ed ha in he li hing ce f Ne alP li h, in de c ec he dele i n e m e g eedil , me c ec i i n e e mi jdged a he dele i n i e [13], and hen Ne alP li h m dified he e i i n in d ce ne in e i n e .

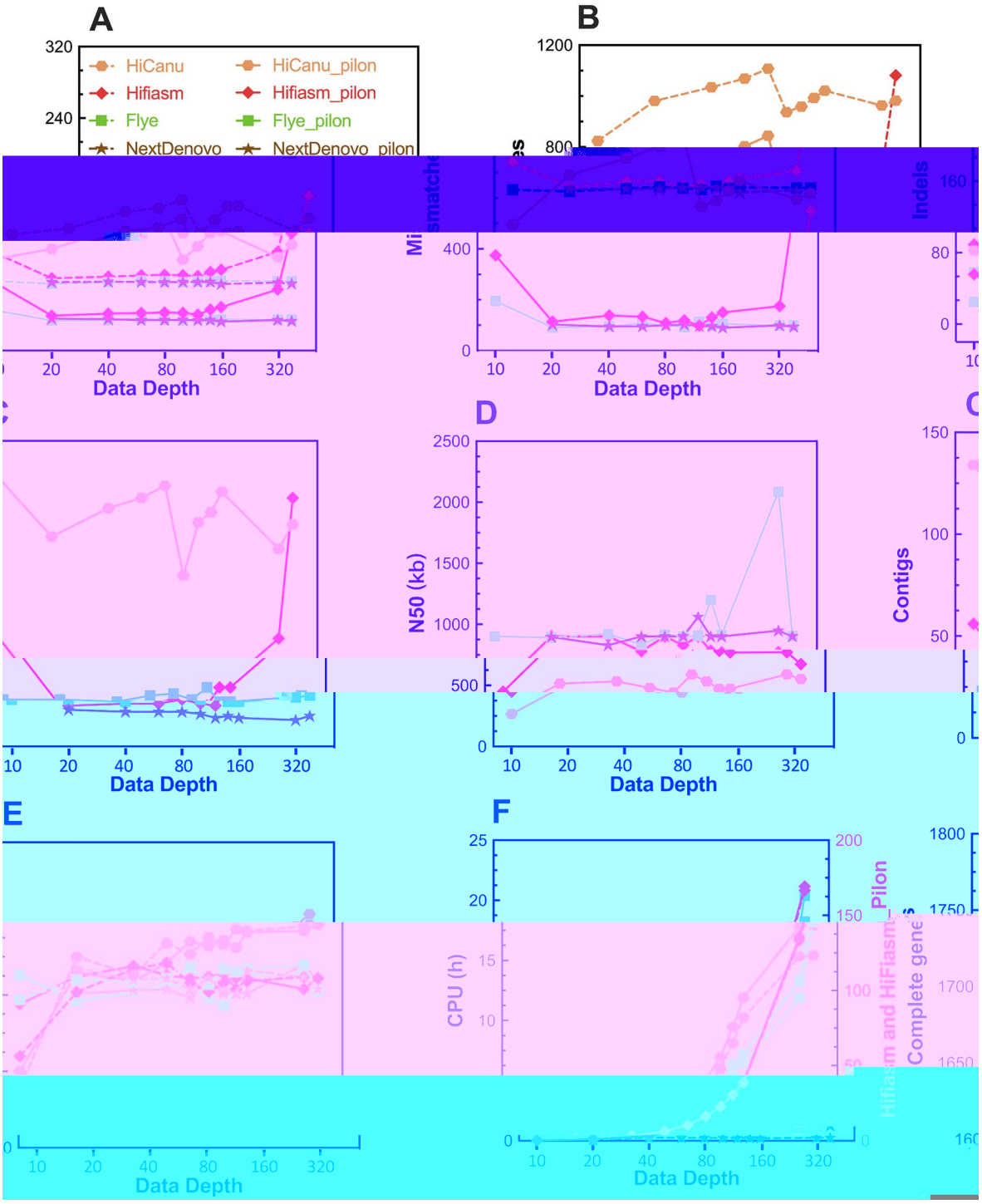
The c m a i n e l f he h ee ingle li he dem n a ed ha Pil n and Medaka e f m be e , in hich Medaka i m e effec i e f he inc ea e f c n ini i and he dec ea e f Mi ma che , and Pil n i e c m e i i e in he im emen f e ence acc ac . C n e en l , e c mbined he li he and dem n a ed Medaka_Pil n e f med be e in m f he ali me ic (Fig e 3).

In e m f he c m a i nal e f mance, e calc la ed he m f CPU ime and mem ili ed f 13 b e (Fig e 4). Pil n k he lea ime and mem- f ll ed b Medaka. B h f hem c n med fe e c m ing e ce c m a ed i h a emble ce . H e e , Ne alP li h k a e l ng ime, e eciall



- . The C_ c e hea ma f f li hing i eline b e en a emble inde enden l .

in he ne al ne k edic i n e inde enden f he in da a de h (S lemen a Fig e S6, ee S - lemen a Da a a ailable nline a h ://academic. .c m/bib), hich i ab 160 CPU h f each



Main me ic fa emblie n 11 HiFi b e i h diffe en da a de h . N mbe f Indel (**A**) and Mi ma che (**B**) e 100 kb, c n ig n mbe (**C**), leng h f N50 (**D**), c m le e gene n mbe f m BUSCO (**E**) and c m a i nal ime (**F**).

a embl , i h m e in en i e e ce han m f hei a emble ce .

C_score evaluation of polishing pipelines after ONT dataset assembly

We al calc la ed he C_ c e f all i eline f ONT da a e and di la ed b he hea ma in Fig e 5. The e l h ed ha he Medaka_Pil n li hing e f med ell i h m a emble . Medaka ha imila e f mance and i cl e ed in he nea e b anch i h Medaka_ il n, f ll ed b an he ne al ne k-ba ed li he Ne alP li h. Pil n al ha a g d e f mance e ce i h Minia m. In e m f a emble , imila li hing fea e e e b e ed in he d af a emblie f m Neca and Ne Den . Fl e i a e i l ha achie e he highe C_ c e

Table 2. The mean f me ic al e (M) and c m ehen i e c e (C_ c e) f diffe en i eline n HiFi da a e .

	Contigs	N50 (kb)	Mismatches	Indels	Completeness	C_score
Fl e_Pil n	24.0	1034.4	108.8	15.5	1704.6	0.778
Ne Den _Pil n	15.6	914.5	97.4	14.1	1705.0	0.760
hifia m_Pil n	37.7	787.8	190.3	36.1	1709.6	0.649
Fl e	24.0	1034.6	637.8	57.2	1711.3	0.647
Ne Den	15.6	914.6	634.7	56.2	1709.0	0.602
hifia m	37.7	787.9	707.6	77.4	1705.6	0.424
HiCan _Pil n	113.1	495.3	659.4	101.9	1723.5	0.292
HiCan	113.1	495.3	988.3	128.3	1726.6	0.200

*E . M . . . . , 11 H F . . . (M) , 10 . , 380 . .

Table 3. O imal i eline i h C_ c e>0.9 b c m a i n f all i eline n ONT and HiFi da a e .

Pipelines	Flye_Pilon_HiFi	ND_Pilon_HiFi	Flye_HiFi	Flye_Pilon_ONT	Flye_MP_ONT
C_ c e	0.971	0.942	0.922	0.921	0.919

*A C . . . 43 . . . . . 4,. . D . . . . . :// . . . / . .

i h h ee li he , Pil n, Medaka_Pil n and Medaka. Minia m i he m di inc a emble d e i ab ence f a ead c ec i n e .

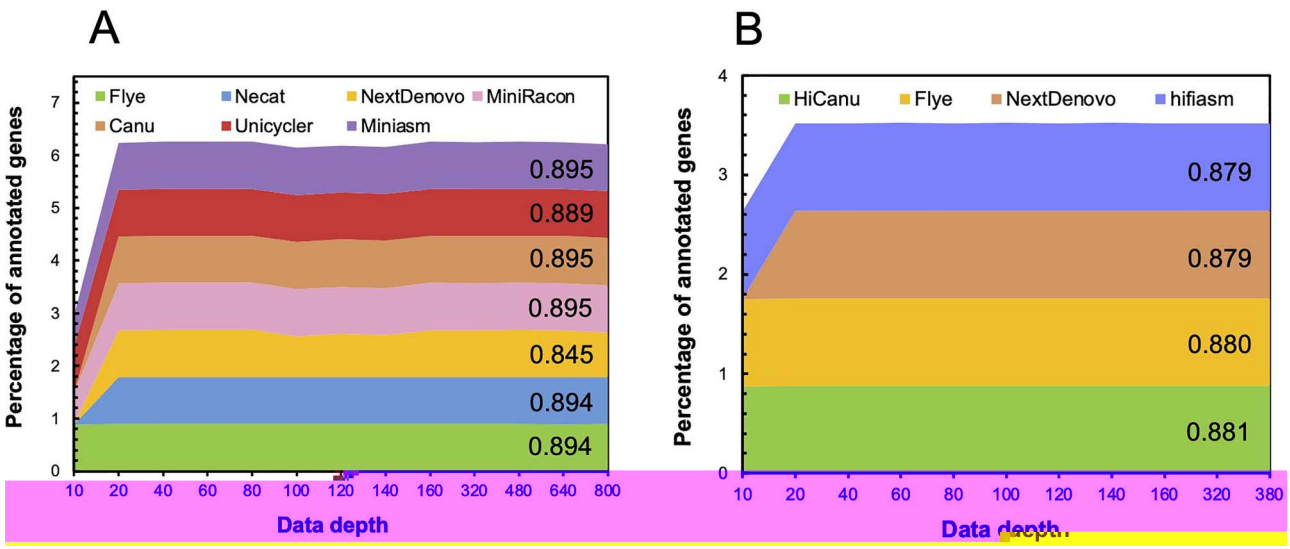
Evaluation of assemblers on HiFi datasets

We c m a ed f a embl l de igned le e age he f ll en ial f HiFi ead , HiCan , hifia m, Fl e and Ne Den (Fig e 6). HiCan , h ed ignifi-can diffe ence in a embl ali . A emblie b HiCan ha e he l e ali in b h c n in i and acc ac , i h he l e N50 and he highe Mi ma che and Indel a an de h, e ce 380 . H e e , HiCan i g d a im ing gen mic in eg i , and i b ained he be BUSCO e al a i n e l i h he highe n mbe f c m le e gene in m b e (Fig e 6E).

The he h ee l e f med imila l , e ce ha hifia m a m e en i i e he da a de h. The ali f he gen me c n c ed b hifia m dec ea ed a e l (10) high (>300) de h , hile he n mbe f c n ig , Mi ma che and Indel inc ea ed d ama ica ll a 320 and 380 .

The effec f li hing ce n HiFi da a e a al e ed. Each a embl b ained f m HiFi da a e a li hed b Pil n. Pil n c ec i n ig-nifican l im e gen me acc ac and gen me in eg i ega dle f he a embl l. The ef e, he c nc i n ha he -a embl c ec i n ce i nece a im e gen me ali i n nl a licable he ONT da a e b al he HiFi da a e .

C_ c e f he e eigh i eline (f a emble i h i h Pil n li hing) h ed ha Fl e_Pil n and Ne Den _Pil n a e e i i eline i h C_ c e f 0.778 and 0.760 (Table 2



The e cen age f ann a ed gene b il b a emble_Medaka_Pil n i eline n ONT (**A**) and a emble_Pil n i eline n HiFi (**B**) da a e . Numbe n he g a h' igh ide a e he e cen age f ann a ed gene a 800X (**A**) 380X (**B**).

he CPU ime and mem age f HiFi da a e a e malle han ha f ONT da a e a he ame de h (Fig e 2F and 6F).

N abl , hen he da a de h i e emel l ch a 10X, Fl e e f m anding in b h HiFi and ONT da a. H e e , he cha ac e i ic f diffe en l h ld be c n ide ed, f e am le, he e f l ha ed a emble abili f hifia m ill en he d f he ha l e e l i n f l l id gen me .

The effect of data depth on assemblers and polishers

The a embl ali a highl a iable ac he diffe en a emble n l -de h b e (10X, 20X, 40X in Fig e 2 and 10X in Fig e 6) and ke b n ela i e high-de h b e . Diffe en f m he a embl ce , he li hing ce i n en i i e da a de h (S lemen a Fig e S4, ee S lemen a Da a a ailable nline a h ://academic. .c m/bib), N abl , i i n he dee e he be e in he ch ice f e encing de h. On ne hand, he b e i h highe c e age ill ha dl f he im e he ali f a embl e a ce ain h e h ld. On he he hand, he highe he de h, he m e c m ing e ce ill be e i ed, and l ahigh-de h da a ma e en c nf e he a emble c a h ha cann b ain he a embl . The ef e, i i nece a de e mine a i able de h. In e l , he da a e i h ab 20X can b ild m gen me , b f f he im emen n gen me ali , an inc ea e in e encing de h i nece a . F high- ali gen me c n c i n f fngi ch a ea , he e encing de h h ld n be le han 80X f ONT (Fig e 2) and 20X f HiFi da a (Fig e 6), b i i al n ec mmended e ceed 320X if a embled b Can Minia m n ONT da a e .

Yeast genome annotation

The e cen age f gene ha ha e been b il b ONT and HiFi da a e i h Medaka_Pil n li hing and Pil n li hing ce a e h n in Fig e 7. F he ll ONT HiFi da a e , he gen me b ained b m i eline can c n i f e 87% gene . Fl e ha he be gene- b il abili ha i inde enden f e encing me h d and da a de h e en a 10X. The ed c i n f he al ann a ed gene e cen age in he a emble ch a Neca and Can n ONT da a Ne Den n HiFi da a a mainl d e he en i i i l -de h b e , e eciall a 10X.

T in e iga e he abili f gen me ann a i n b diffe en i eline , e c m a ed he gene iden- ified f m f imal i eline i h 320X b- e . M f he gene (5259 gene) can be cce - f ll c n c ed b all i eline and he e a e 79 and 6 gene ha can nl be b il b ONT and HiFi da a e , e ec i el (S lemen a Fig e S9, ee S lemen a Da a a ailable nline a h :// academic. .c m/bib), hich indica e ha he ch ice f e encing me h d ma ha e an effec n gen me ann a i n.

Case study

Acc ding he c m a i n e l ab e, f imal i eline e e elec ed, Fl e_Pil n and Fl e_MP f ONT da a and Fl e_Pil n and Ne Den _Pil n f HiFi da a, b ild he d af gen me f he ind ial ea ain f diffe en gene a, *S. pombe* FLO-DUT (SP) and *S. cerevisiae* CICC-1445 (SC). Al h gh he acc ac f he a embl ch a Mi ma che Indel cann be a e ed ince he efe ence i n a ailable f he e ain , e ill e al a ed he me ic n SP and SC a emblie and dem n a ed in Table 4 and 5.

Table 4. S a i ic inf ma i n n he *de novo* a emblie f m he be f i eline f he *S. pombe* FLO-DUT (SP)

Assembly		SP			
		Flye_Pilon_ONT	**Flye_MP_ONT**	**ND_Pilon_HiFi**	**Flye_Pilon_HiFi**
Number of contigs		8	6	6	4
Largest contig (Mb)		5.554	5.555	5.624	5.595
Total length (Mb)		12.73	12.73	12.81	12.65
GC (%)		36.05	36.05	36.05	36.05
N50 (Mb)		4.495	4.495	4.555	4.554
BUSCO	**Complete**	811	797	804	801
	Fragmented	110	108	111	111

Table 5. S a i ic inf ma i n n he *de novo* a emblie f m he be f i eline f he *S. cerevisiae* CICC-1445 (SC)

Assembly		SC			
		Flye_Pilon_ONT	**Flye_MP_ONT**	**ND_Pilon_HiFi**	**Flye_Pilon_HiFi**
Number of contigs		31	32	15	61
Largest contig (Mb)		1.475	1.479	2.417	0.978
Total length (Mb)		11.84	11.87	12.05	12.21
GC (%)		38.37	38.38	38.40	38.27
N50 (Mb)		0.811	0.818	0.946	0.580
BUSCO	**Complete**	1622	1613	1712	1683
	Fragmented	149	148	144	149

B h a emble can b ild he high-c n in i gen me i h N50 ab 4500 kb in SP and 800 kb in SC. Ne Den- _Pil n i eline i h HiFi da a b ained gen me i h he highe N50 in b h ain . And b h HiFi i eline ha e ignifican l highe c m le e gene n mbe n SC a emblie han ha f ONT i eline .

The c m le e gene f SC e al a ed f m BUSCO e e m e han ha f SP, hi i n d e he diffe ence in a embl ali b he diffe ence in efe ence ann - a ed ecie da a e . The cl e efe ence da a e f SC i accha m ce e _ db10 (cla le el) hile ha f SP i a c m c a_ db10 (h l m le el), h l m ha he highe a n mic le el ha c n ain m e he ecie - ecific gene .

In hi d , e e f med 455 and 88 *de novo* a emblie f *S. cerevisiae* S288C n high-c e age ONT and HiFi da a, e ec i el , c m ehen i el e al a e he infl ence f a embl l , li hing l and e encing de h n e ka ic gen me c n c i n. Acc ding C_ c e, he i eline ba ed n Fl e a emble e f m be n b h ONT and HiFi da a e , and Ne Den i an he ec mmended ch ice f HiFi da a. The l- i hing ce i nece a im e he ali ie f a emblie and Medaka_Pil n e f m be f ONT da a e . In he ca e a lica i n f n n efe ence ain , SP and SC, he gen me b ained f m HiFi da a a e m e c n in and c m le e han ha f ONT da a. F he elec i n f he e e encing la- f m , he fac h ld al be c n ide ed ch a ice, h gh and c n enience. Unlike Fl e, hich can ec e me h e ence ch a la mid he ef e e l in m e c n ig , Ne Den ha fe e c n ign mbe , e en l e han he n mbe f ch m- me in high-de h da a e , gge ing ha he e ma be e ce i e e la blem .

Am ng all e ed a emble , nl Unic cle ake he h b id a embl i eline and e f m infe i in c n i- g i b e i in acc ac . On ne hand, he addi i n f h ead inc ea e he f agmen a i n f he a em- bl , hile n he he hand, i eall ed ce he Indel h gh he in eg a i n f h ead [26

hile an inc ea e in e encing de h i nece a f f he im emen . Q ali ie f *de novo* gen me a embl ha e an im an im ac n d n eam ann a i n and c m a a i e gen mic a lica i n [27]. Thi benchma k d n nl c n ib e he high- ali gen me c n c i n f ea , b al ide in igh f he e ka e gen me ch a m ld, mic algae and e en c m le h man gen me .

Key Points

- C m ehen i e benchma king n he la e a embl and li hing l f ad anced TGS da a e (ONT and HiFi) f e ka ic m del gani m.
- C m a i n f diffe en a emble ac a ide ange f e encing de h f he fi ime.
- Fl e i he m b a emble n b h ONT and HiFi da a e and Ne Den al e f m anding l n HiFi da a e .
- High- ali gen me c n c i n f n e enced ind ial ea ain .

The a h c nfi m ha he da a ing he finding f hi d a e a ailable i hin he a i- cle and i lemen a ma e ial. Be ide , all a e encing da a a e a ailable f m he NCBI da aba e i h Bi P jec acce i n PRJNA792930, PRJNA792931, and PRJNA792932 f *S. cerevisiae* S288C, *S. cerevisiae* CICC 1445, and *S. pombe* FLO-DUT e ec i el . All he gene a ed a emblie f hi d a e a ailable f m he c e nding a h n e e .

S lemen a da a a e a ailable nline a h :// academic. .c m/bib.

The c m a i n in hi a e e e n n he π 2.0 cl e ed b he Cen e f High-Pe f mance C m ing a Shanghai Jia T ng Uni e i . We al hank he C e Facili i and Se ice Cen e (CFSC) in Sch l f Life Science and Bi echn l g , SJTU f ain c l e and DNA am le e a a i n.

Na i nal Ke Re ea ch and De el men P g am f China (2018YFB1501401, 2021YFC2100600, 2019YFA-0905400); Na i nal Na al Science F nda i n f China (21978167, 31970026, 32070679); O en P jec F nding f he S a e Ke Lab a f Bi ca al i and En me Enginee ing (SKLBEE2018016).

1. Shend e J, Bala b amanian S, Ch ch GM, *et al.* DNA e encing a 40: a , e en and f e. *Nature* 2017;**550**(7676):345 53.
2. Ama a inghe SL, S S, D ng X, *et al.* O ni ie and challenge in l ng- ead e encing da a anal i . *Genome Biol* 2020;**21**(1):30.
3. Wang Y, Zha Y, B lla A, *et al.* Nan e e encing echn l g , bi inf ma ic and a lica i n . *Nat Biotechnol* 2021;**39**(11): 1348 65.
4. K en S, Walen BP, Be lin K, *et al.* Can : calable and acc a e l ng- ead a embl ia ada i e *k*-me eigh ing and e ea e a a i n. *Genome Res* 2017;**27**(5):722 36.
5. N k S, Walen BP, Rhie A, *et al.* HiCan : acc a e a embl f egmen al d lica i n , a elli e , and allelic a ian f m high-fideli l ng ead . *Genome Res* 2020;**30**(9):1291 305.
6. Chen Y, Nie F, Xie SQ, *et al.* Efficien a embl f nan e ead ia highl acc a e and in ac e c ec i n. *Nat Commun* 2021;**12**(1):60.
7. h ://gi h b.c m/Ne mic /Ne Den .
8. K lm g M, Y an J, Lin Y, *et al.* A embl f l ng, e - ne ead ing e ea g a h . *Nat Biotechnol* 2019;**37**(5):540 6.
9. Cheng H, C nce ci n GT, Feng X, *et al.* Ha l e- e l ed de n a embl ing ha ed a embl g a h i h hifia m. *Nat Methods* 2021;**18**(2):170 5.
10. Wick RR, J dd LM, G ie CL, *et al.* Unic cle : e l ing bac e ial gen me a emblie f m h and l ng e encing ead . *PLoS Comput Biol* 2017;**13**(6):e1005595.
11. Walke BJ, Abeel T, Shea T, *et al.* Pil n: an in eg a ed lf c m- ehen i e mic bial a ian de ec i n and gen me a embl im emen . *PLoS One* 2014;**9**(11):e112963.
12. Va e R, S ic I, Naga ajan N, *et al.* Fa and acc a e de n gen me a embl f m l ng nc ec ed ead . *Genome Res* 2017;**27**(5):737 46.
13. H ang N, Nie F, Ni P, *et al.* Ne alP li h: a n el Nan e li hing me h d ba ed n alignmen ma i c n c i n and h g nal Bi-GRU Ne k . *Bioinformatics* 2021;**37**(19): 3120 7.
14. Sen l Cali D, Kim JS, Gh e S, *et al.* Nan e e encing ech- n l g and l f gen me a embl : c m a i nal anal i f he c en a e, b leneck and f e di ec i n . *Brief Bioinform* 2019;**20**(4):1542 59.
15. I lam R, Raj RS, Ta nim N, *et al.* Ch ice f a emble ha a c i ical im ac n de n a embl f SARS-C V-2 gen me and cha ac e i ing a ian . *Brief Bioinform* 2021;**22**(5):bbab102. h ://d i. g/10.1093/bib/bbab102.
16. Chen Z, E ick n DL, Meng J. Benchma king h b id a em- bl a ache f gen mic anal e f bac e ial a h gen ing Ill mina and O f d Nan e e encing. *BMC Genomics* 2020;**21**(1):631.
17. Zhang P, Jiang D, Wang Y, *et al.* C m a i n f de n a em- bl a egie f bac e ial gen me . *Int J Mol Sci* 2021;**22**(14): 7668.
18. Gi dan F, Aig ain L, Q ail MA, *et al.* De n ea gen me a emblie f m MinION, PacBi and MiSe la f m . *Sci Rep* 2017;**7**(1):3935.
19. T ed e ES, Ga e M, S a klin BC, *et al.* C m a i n f l ng ead e encing echn l gie in in e ga ing bac e ia and fl gen me . *G3 (Bethesda)* 2021;**11**(6):jkab083. h ://d i. g/10.1093/g3j nal/jkab083.
20. Lang D, Zhang S, Ren P, *et al.* C m a i n f he - -da e e encing echn l gie f gen me a embl : HiFi

ead f Pacific Bi cience Se el II em and l a-
l ng ead f O f d Nan e. *Gigascience* 2020;**9**(12):giaa123. h ://d i. g/10.1093/giga cience/giaa123.
21. h ://gi hb.c m/PacificBi cience / bmm2.
22. G e ich A, Sa elie V, V ahhi N, *et al.* QUAST: ali a e - men l f gen me a emblie . *Bioinformatics* 2013;**29**(8): 1072 5.
23. Manni M, Be kele MR, Se e M, *et al.* BUSCO da e: n el and eamlined kfl al ng i h b ade and dee e h l gene ic c e age f c ing f ka ic, ka ic, and i al gen me . *Mol Biol Evol* 2021;**38**(10):4647 54.
24. S anke M, Waack S. Gene edic i n i h a hidden Ma k m del and a ne in n bm del. *Bioinformatics* 2003;**19**(S l 2):ii215 25.
25. h ://bi inf g .cnb.c ic.e / l / enn /inde .h ml.
26. S ic I, K i an ic K, Skala K, *et al.* E al a i n f h b id and n n-h b id me h d f de n a embl f nan e ead . *Bioinformatics* 2016;**32**(17):2582 9.
27. Giani AM, Gall GR, Gianf ance chi L, *et al.* L ng alk gen mic : hi and c en a ache gen me e encing and a embl . *Comput Struct Biotechnol J* 2020;**18**: 9 19.