The first gapless, reference-quality, fully annotated genome from a Southern Han Chinese individual

Kuan-Hao Chao[1,2], Aleksey V Zimin[2,3], Mihaela Pertea[2,3], Steven L Salzberg[1,2,3,4]

[1] Department of Computer Science, Johns Hopkins University, [2] Center for Computational Biology, Johns Hopkins University, [3] Department of Biomedical Engineering, Johns Hopkins University, [4] Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21211, USA

https://github.com/JHUCCB/ChineseHanSouthGenome

https://doi.org/10.1101/2022.08.08.503226

https://www.ncbi.nlm.nih.gov/assembly/GCA_024586135.1/

https://khchao.com

Introduction

- Used **Hi-Fi long reads** to assemble the genome of a Southern Han Chinese male, ***Han1***
- Filled in gaps using T2T-CHM13 as a guide. ***Han1* is gap-free**
- Han1 contains **3,099,707,698 bases**
- Annotated Han1 using Liftoff: **60,708 putative genes** (20,003 protein coding genes)
- The **first gene-level comparison** between two finished, annotated individual human genomes

Genome Assembly

Sequence data

Sequence type	Basecaller	Number of reads	N50 read length (bp)	Mean read length (bp)	Maximum length	Total length (Gbp)	Genome coverage
PacBio HiFi	ccs v4.0.0	5,570,675	21,499	21,989	50,388	122.50	39.45x
ONT Ultralong	Guppy v4.0.11	3,110,293 (9.11% > 100 Kb)	83,822	34,136	2,495,296	106.17	34.75x

Draft HiFi and ONT Assemblies

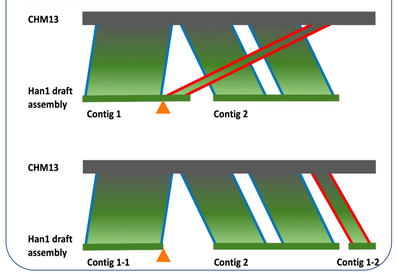
Assembler	Sequencing data	Assembled sequence (bp)	Contig N50	Number of contigs	Quality value
Hifiasm v0.16.1-r375	PacBio HiFi	3,110,501,483	95,769,069	182	57.8 [1]
Flye v2.5	ONT Ultralong	2,974,205,132	40,850,737	1,658	25.6 [1]

Han1 Assembly

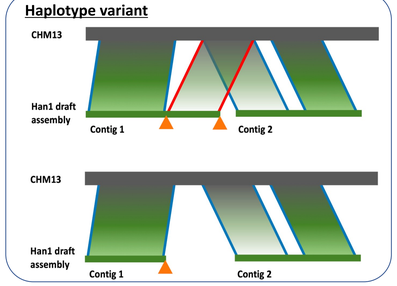
- **Primary:** Hifiasm assembly
- **Secondary:** Flye Assembly
- **Reference:** T2T-CHM13
- **Solving misassembled contigs.**
- **Closing Gaps** on primary assembly
- 1. HiFi reads 2. Fly assembly 3. T2T-CHM13

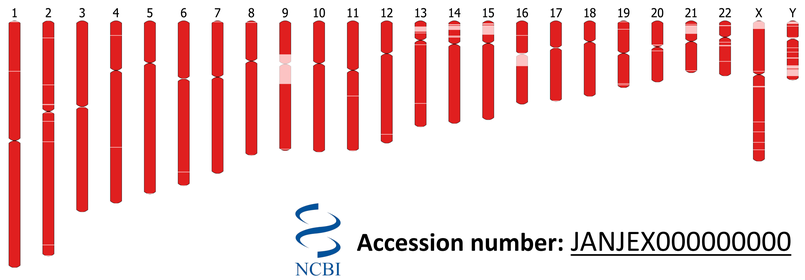
Chromosome	Han1 total (bp)	Non-HG00621 sequence (inserted from CHM13) (bp)	Ratio of source sequences (HG00621:CHM13)
1	249,525,787	119,184	0.9995
2	242,739,747	2,482,037	0.9898
3	200,211,729	377,991	0.9981
4	192,045,028	518,393	0.9973
5	181,667,637	494,129	0.9973
6	170,861,069	314,798	0.9982
7	160,865,769	107,243	0.9993
8	145,880,131	791,768	0.9946
9	148,018,047	35,504,706	0.7601
10	135,316,043	585,347	0.9957
11	135,129,219	874,841	0.9935
12	134,132,185	102,971	0.9992
13	111,903,191	10,782,722	0.9036
14	101,435,482	5,090,291	0.9498
15	101,210,777	12,429,469	0.8772
16	95,412,483	13,280,238	0.8608
17	83,450,189	1,080,955	0.9870
18	78,996,361	210,798	0.9973
19	61,978,944	1,089,081	0.9824
20	65,189,243	963,587	0.9852
21	45,827,290	5,613,897	0.8775
22	50,610,422	5,397,082	0.8934
X	154,227,164	7,056,525	0.9542
Y	53,057,190	14,607,629	0.7247
M	16,571	0	1.0000
Total	3,099,707,698	119,875,682	0.9614

Wrong-order misassembly



Haplotype variant





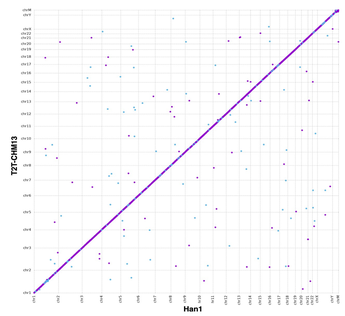
NCBI **Accession number:** JANJEX000000000

Comparing *Han1* with Published Assemblies

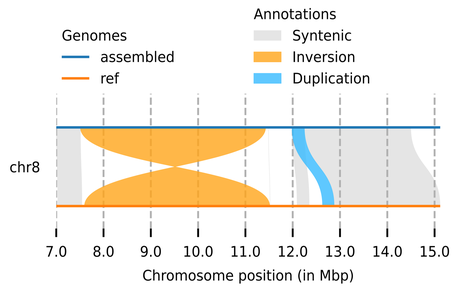
Genome	Ethnicity	Contig N50 (Mb)	Number of contigs	Number of gaps	Assembly size (Gb)
Han1[1]	Southern Han Chinese	148.02	25	0	3.10
HG00621 (hifiasm)[2]	Southern Han Chinese	95.77	182	157	3.11
T2T-CHM13v2.0 [3]	Northern European	150.62	25	0	3.12
HJ-H1[4]	Northern Han Chinese	28.15	1,330	427	3.07
HJ-H2 [4]	Northern Han Chinese	25.90	896	390	2.91
NH1 [4]	Northern Han Chinese	3.60	11,019	8,484	2.89
HX1 [4]	Southern Han Chinese	8.33	5,843	4,025	2.93
YH2.0 [5]	Southern Han Chinese	0.02	361,157	235,514	2.91
TJ1.p0 [6]	Tujia	13.67	1430	907	2.87
TJ1.p1 [6]	Tujia	13.70	1426	873	2.87
ZF1 [7]	Tibetan	23.62	1384	1360	2.85
GRCh38.p14 [8]	Mixed	57.88	994	804	3.10

Han1 / T2T-CHM13 Assembly Comparison

- Collinearity (MUMmer4)



- Inversion on β defensin gene cluster (Plotsr)



Gene Annotation

GRCh38 / T2T-CHM13 / *Han1* Gene Count Comparison

- Two-pass annotation lift-over from T2T-CHM13 to Han1 using Liftoff

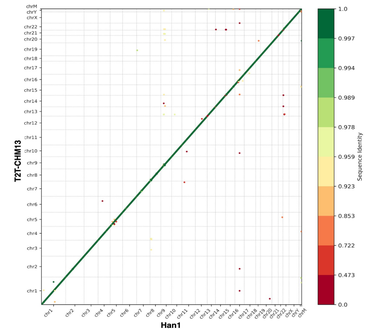
Gene biotype	Count in GRCh38	Count in T2T-CHM13	Count in Han1
protein coding	19,871	20,022	20,003
lncRNA	17,793	18,389	18,321
pseudogene	15,357	16,027	15,881
miRNA	1,914	2,046	2,058
transcribed pseudogene	1,221	1,262	1,232
rRNA	38	765	658
other	2,506	2,629	2,556
Total	58,700	61,140	60,708

Han1 / T2T-CHM13 Gene Content Comparison

- **Liftofftools**: [1] Variant

Mapped genes with CDS features								
Conserved genes (82%)		Altered genes (18%)						
Identical	Synonymous	Nonsynonymous	Start lost	Stop gained	Truncated	Frameshift	In-frame insertion	In-frame deletion
14130	2442	3200	21	31	1+39 (5′+3′)	143	110	92

- **Liftofftools**: [2] Synteny



- **Liftofftools**: [3] Clusters

Copy number in T2T-CHM13	Genes in T2T-CHM13	Copy number in Han1
5	IGHVIII	3
7	TBC1D3	4
9	AMY	7
10	SPDYE	7
34	FAM90A	16

Homozygous Altered Genes

Gene name	RefSeq protein length	Protein length		Protein length ratio (vs RefSeq)		Amino acid identity score (vs RefSeq)	
		Han1	T2T-CHM13	Han1	T2T-CHM13	Han1	T2T-CHM13
MUC19	6,985	1,241	5,332	0.178	0.763	0.278	0.943
AQP12A	295	163	295	0.553	1.000	0.551	0.986
RETNLB	111	111	14	1.000	0.126	1.00	0.295
TCP11X1	407	407	201	1.000	0.494	0.997	0.417
DEFB126	111	111	82	1.000	0.739	0.936	0.576
TPSB2	275	275	166	1.000	0.604	0.927	0.584
PBOV1	135	136	135	1.010	1.000	1.000	0.875
GOLGA6L10	536	550	522	1.026	0.974	0.945	1.000
KLHDC7B	1,235	1,211	1,215	0.981	0.984	0.978	0.981
NBPF19	3,843	3,283	3,772	0.854	0.982	0.927	0.961
RP1L1	2,400	2,417	2,464	1.007	1.027	0.989	0.970
TMEM82	343	343	344	1.000	1.000	1.000	0.994
KIR2DL3	341	341	341	1.00	1.000	1.000	0.956

Conclusions

- **First gapless** Southern Han Chinese Genome
- **First fully annotated** genome
- **First gene content comparison** between two individuals

Contact

Kuan-Hao Chao : kh.chao@cs.jhu.edu; https://khchao.com/
Steven Salzberg : salzberg@jhu.edu; https://salzberg-lab.org/
Mihaela Pertea : mpertea@jhu.edu; http://www.ccb.jhu.edu/people/mpertea/