National Cancer Institute

Helicobacter pylori Genome Project (HpGP)

Difei Wang, Ph.D.

Bioinformatics Manager
Research Analysis Support Group
Cancer Genomics Research Laboratory
Frederick National Laboratory for Cancer Research
Leidos Biomedical Research, Inc.

2024-02-21

Supporting the Division of Cancer Epidemiology and Genetics

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

CGR Cancer Genomics Research

H. pylori is a carcinogen

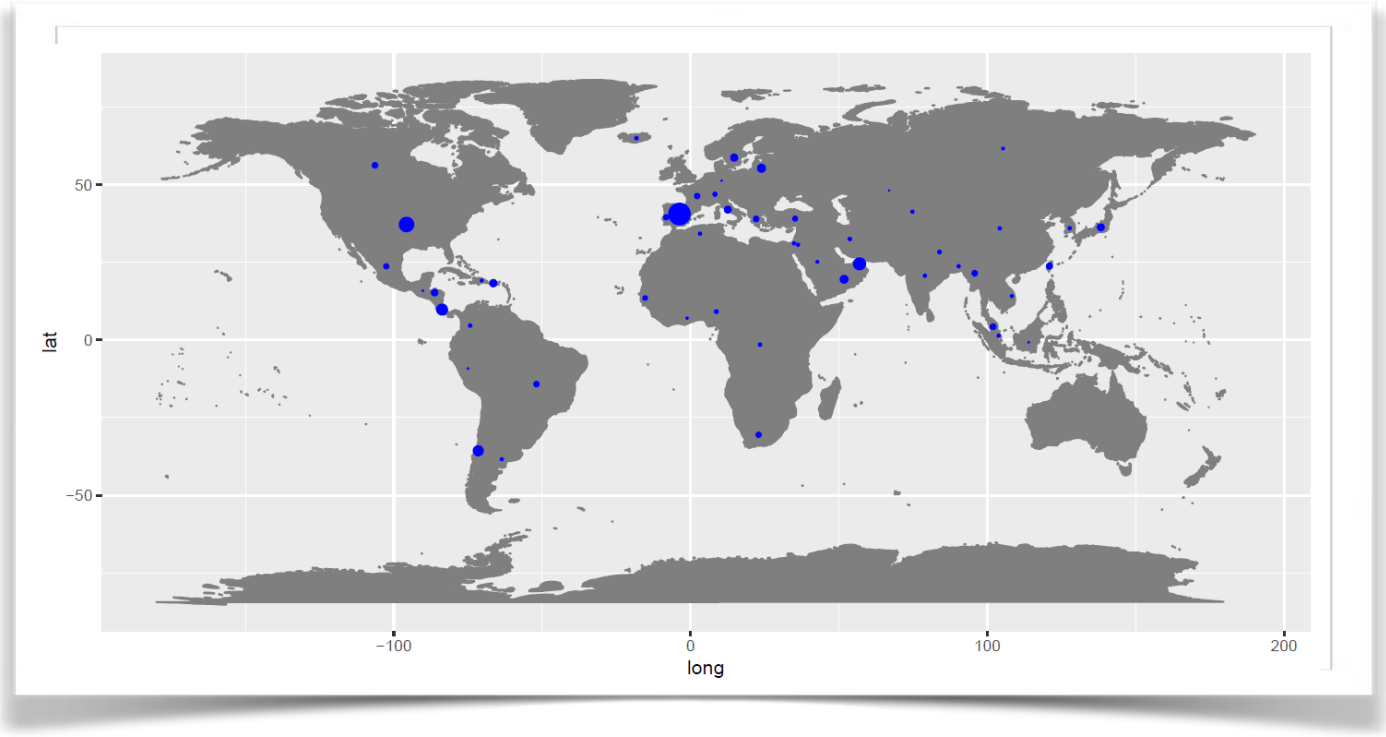
- H. pylori is a gram-negative bacterium and has been linked with stomach cancer. About 2 out of 3 adults worldwide are infected with *H. pylori*.
- Still limited recognized virulence factors (e.g. *cagA* and *vacA*, highly prevalent in high-risk areas, e.g. East Asia)
- Genome first sequenced in 1997. Genome size is about 1.6 Mb. About 1,500 coding genes.
- A few hundred complete genomes in GenBank in 2016. Most of them from RSII or Illumina short-reads.
  - Still a limited number of genomes per geographic site
- Genome-wide base modifications first profiled in 2014 ("Methylome")
- Newly available technology
  - Single molecule, real-time (SMRT) sequencing: PacBio RSII, Sequel and Sequel II and Nanopore platforms.

HpGP was launched in 2017.

Co-PIs: Dr. Constanza Camargo
Dr. Charles Rabkin

https://www.cancer.gov/about-cancer/causes-prevention/risk/infectious-agents/h-pylori-fact-sheet#r23

~ 51 HpGP Collaborating Centers



Each Center to Contribute:
Gastric Cancer (GC)
Adv. Intestinal Metaplasia (IM)
Non-atrophic Gastritis (NAG) and others

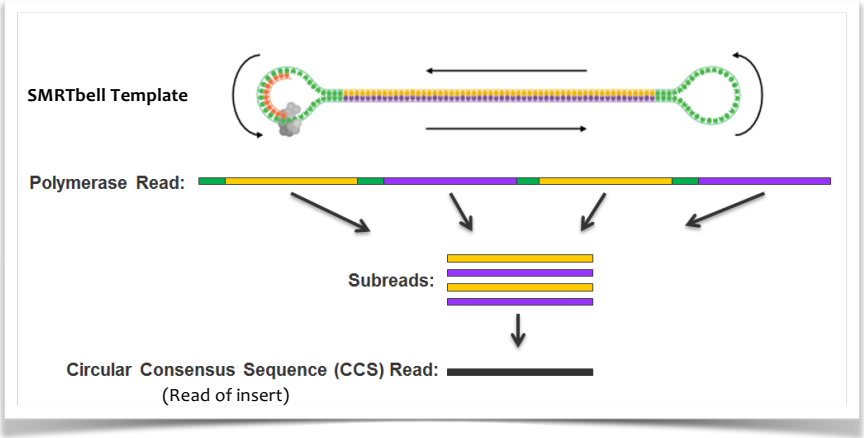
Total more than 1033 samples collected and sequenced so far.

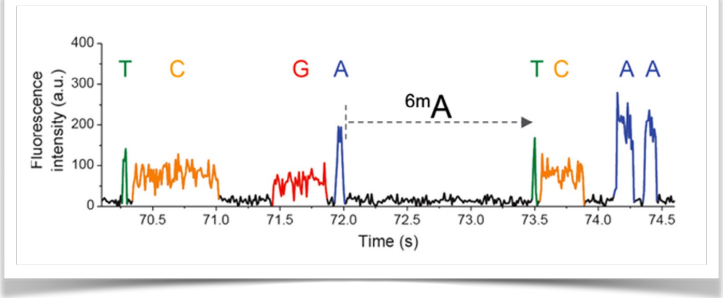
HpGP Objectives

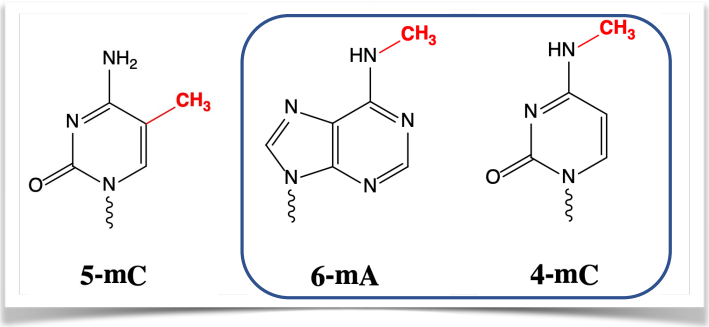
- To map the global population structure of > 1000 *de novo* assembled complete *H. pylori* genomes
- To characterize the spectrum of genomic and epigenomic variations of benign strains of *H. pylori*
- To identify molecular features that may contribute to pathologic effects
- To establish a resource repository of multidimensional data and well-characterized strains for utilization by the scientific community

Co-PIs: Dr. Constanza Camargo
Dr. Charles Rabkin

Single Molecule, Real-Time (SMRT) Sequencing







https://www.pacb.com/

A Pilot Study in 2017

- Mexico: 43 [23 Cancer, 17 Gastritis, 3 Others]
- Honduras: 27 [8 Cancer, 9 NAG, 10 IM]
- **Colombia: 26 [2 Cancer, 12 IM, 12 NAG]**
- Lativa: 21 [1 Cancer, 10 NAG, 10 IM]
- Taiwan: 20 [10 Cancer, 10 NAG]

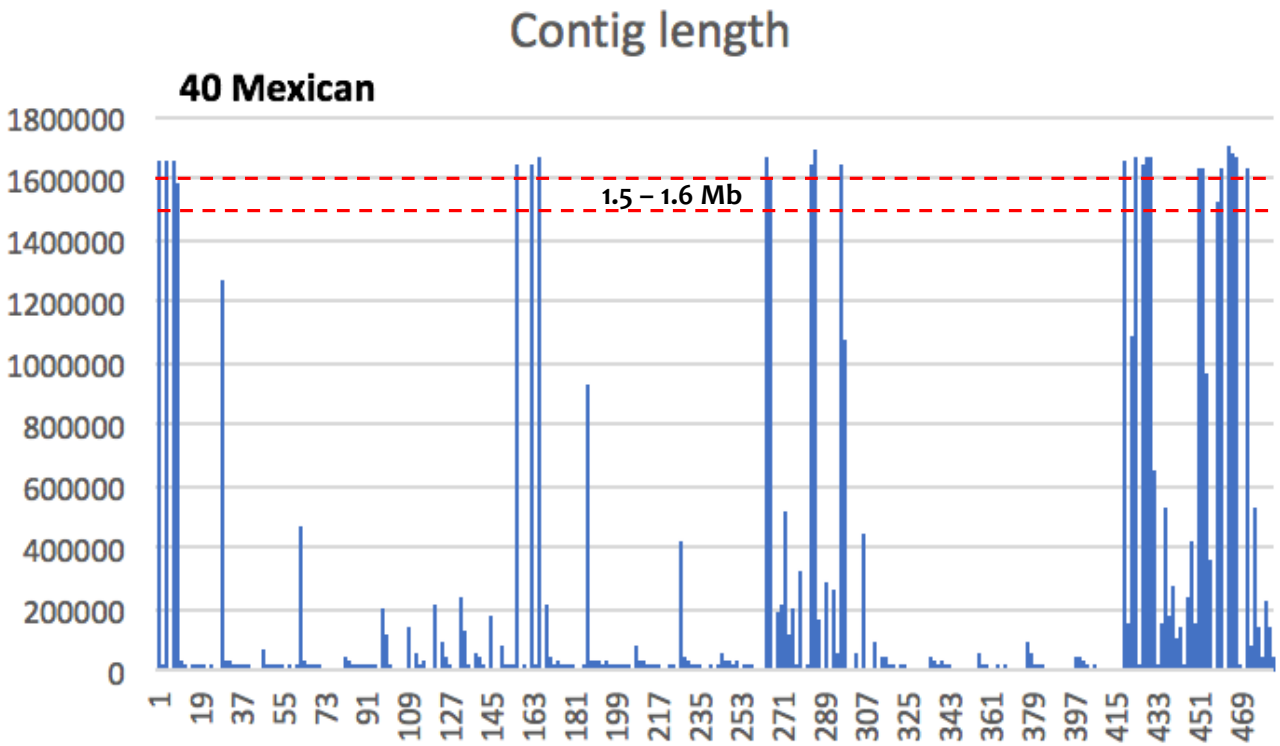
- Greece: 8
- Peru: 10

PacBio RSII

~ 155 *H. pylori* genome sequences were sequenced on PacBio RSII, de novo assembled and delivered. (Frederick)

In order to know how well they were assembled, we need to do some quality checking and align the assembled genome sequences to the know sequences.

Number and Length of Contigs



~ 40%

With more than 1 contig and each of them much less than 1.5Mb

Known *H. pylori* Genomes

ANCIENT MICROBIOME

The 5300-year-old *Helicobacter pylori* genome of the Iceman

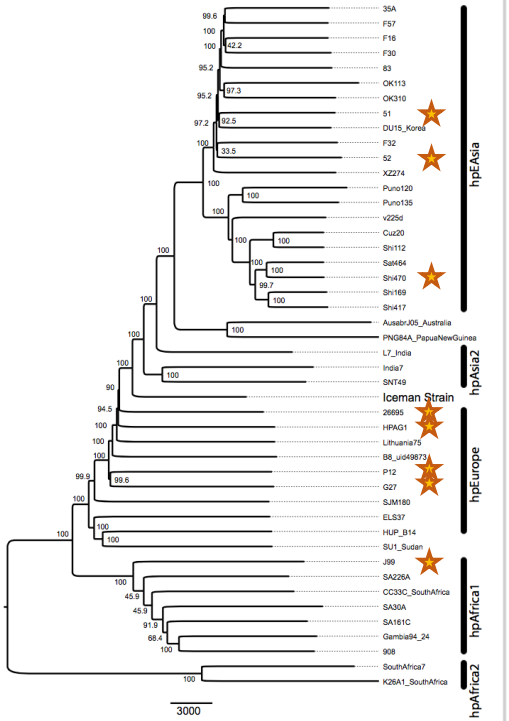
Frank Maixner,[1]*† Ben Krause-Kyora,[2]† Dmitrij Turaev,[3]† Alexander Herbig,[4,5]† Michael R. Hoopmann,[6] Janice L. Hallows,[6] Ulrike Kusebauch,[6] Eduard Egarter Vigl,[7] Peter Malfertheiner,[8] Francis Megraud,[9] Niall O'Sullivan,[1] Giovanna Cipollini,[1] Valentina Coia,[1] Marco Samadelli,[1] Lars Engstrand,[10] Bodo Linz,[11] Robert L. Moritz,[6] Rudolf Grimm,[12] Johannes Krause,[4,5]‡ Almut Nebel,[2]‡ Yoshan Moodley,[13,14]‡ Thomas Rattei,[3]‡ Albert Zink[1]*‡

The stomach bacterium *Helicobacter pylori* is one of the most prevalent human pathogens. It has dispersed globally with its human host, resulting in a distinct phylogeographic pattern that can be used to reconstruct both recent and ancient human migrations. The extant European population of *H. pylori* is known to be a hybrid between Asian and African bacteria, but there exist different hypotheses about when and where the hybridization took place, reflecting the complex demographic history of Europeans. Here, we present a 5300-year-old *H. pylori* genome from a European Copper Age glacier mummy. The "Iceman" *H. pylori* is a nearly pure representative of the bacterial population of Asian origin that existed in Europe before hybridization, suggesting that the African population arrived in Europe within the past few thousand years.

Oetzi found in Italy in 1991

hpAsia2 (Indian & Europe), cagA+, vacA+

Selected 9 complete *H. pylori* genomes from NCBI for alignment.



Maixner, F. *et al* Science 351, 162 (2016)

Genome Alignment of 9 Published *H. pylori* Genomes in GenBank

hpEurope
26695

G27

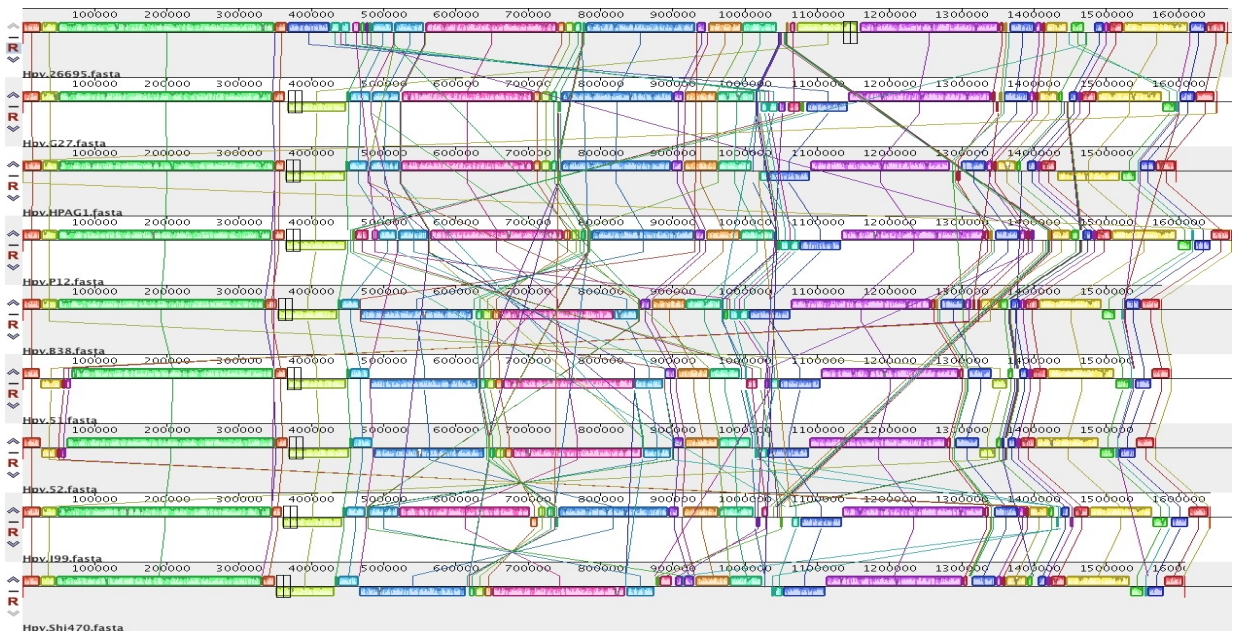
HPAG1

P12

B38

hspEAsia
51

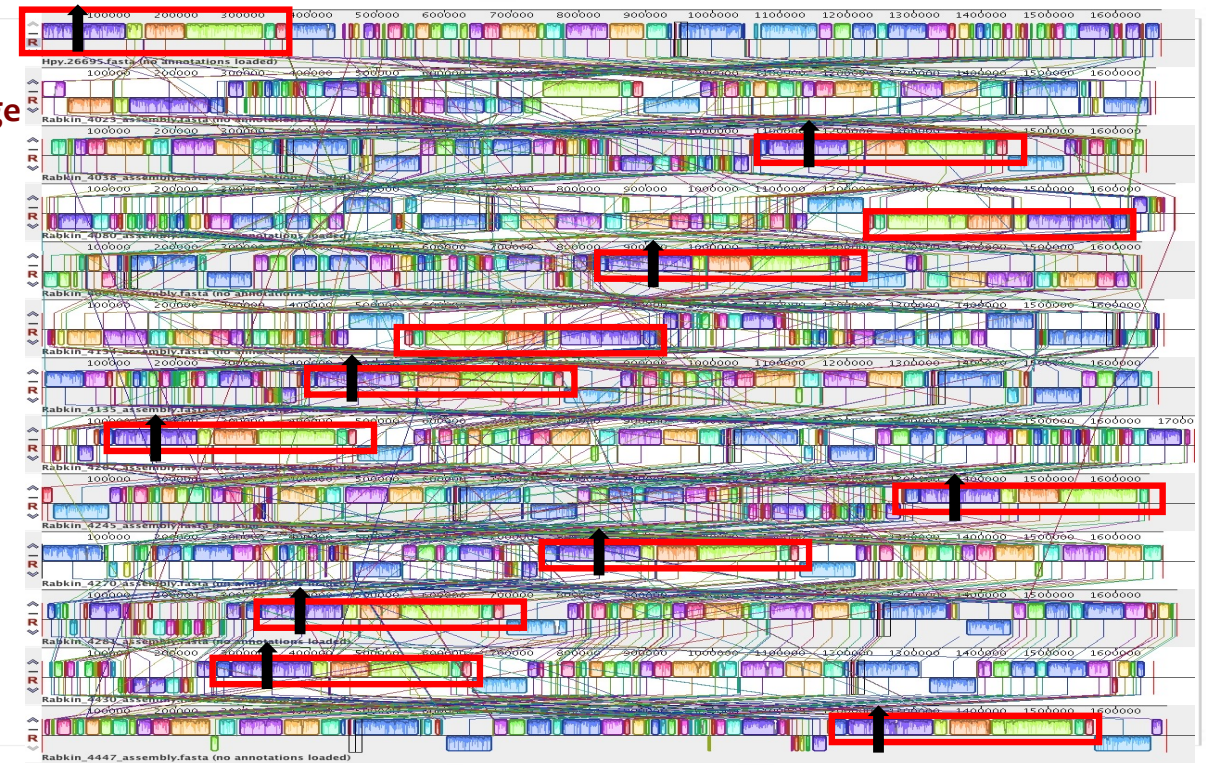
52
hspWAfrica
J99
hspAmerInd
Shi470



Genome Alignment of 12 Colombian (NAG) *H. pylori* Genomes to HP26695

hpEurope 26695

Colombia_NoChange
Rabkin_4023
….
Rabkin_4447



Genome Alignment of 12 Colombian (Progression) *H. pylori* Genomes to HP26695

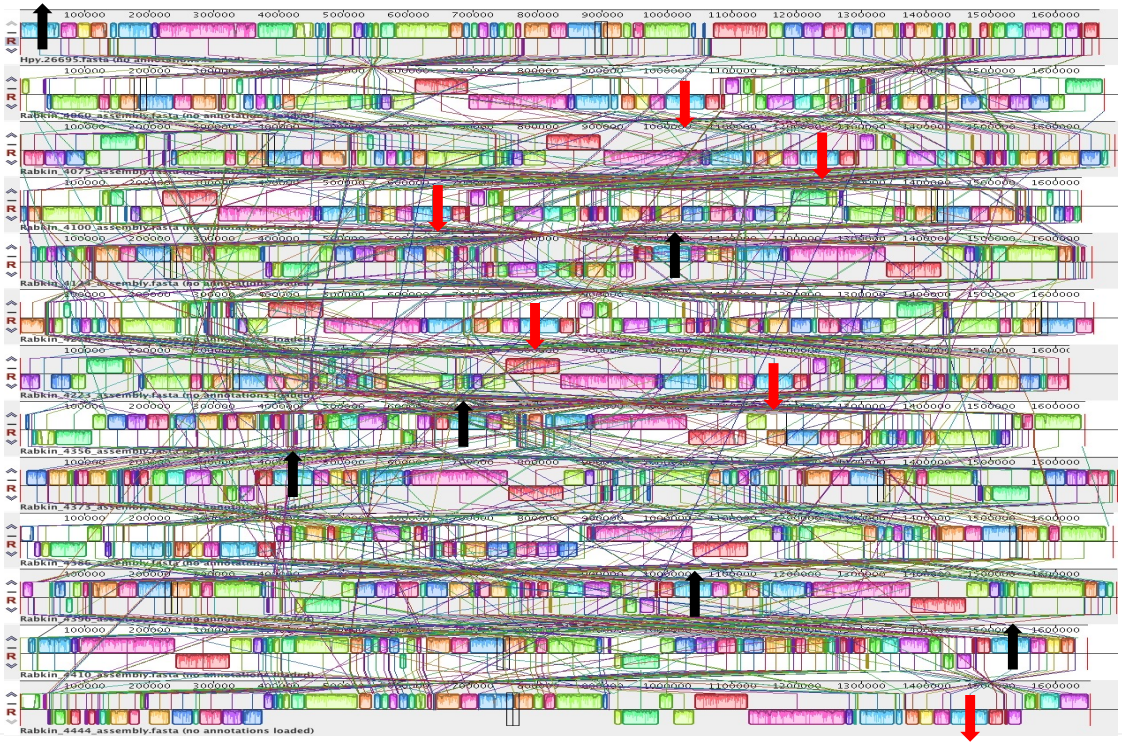
hpEurope 26695

Colombia_Progression

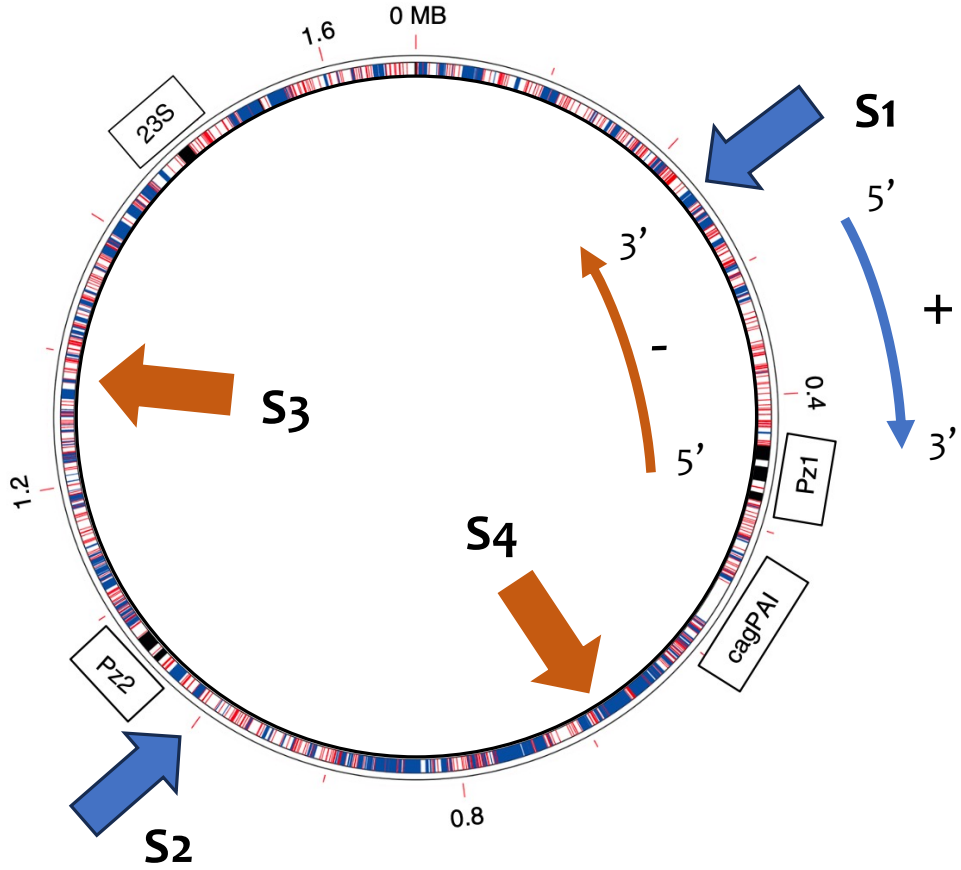
Rabkin_4060

….

Rabkin_4444



What Happened?



H. Pylori genome is circular.

De novo assembling gives you the assembled sequence with inconsistent starting sites/points.

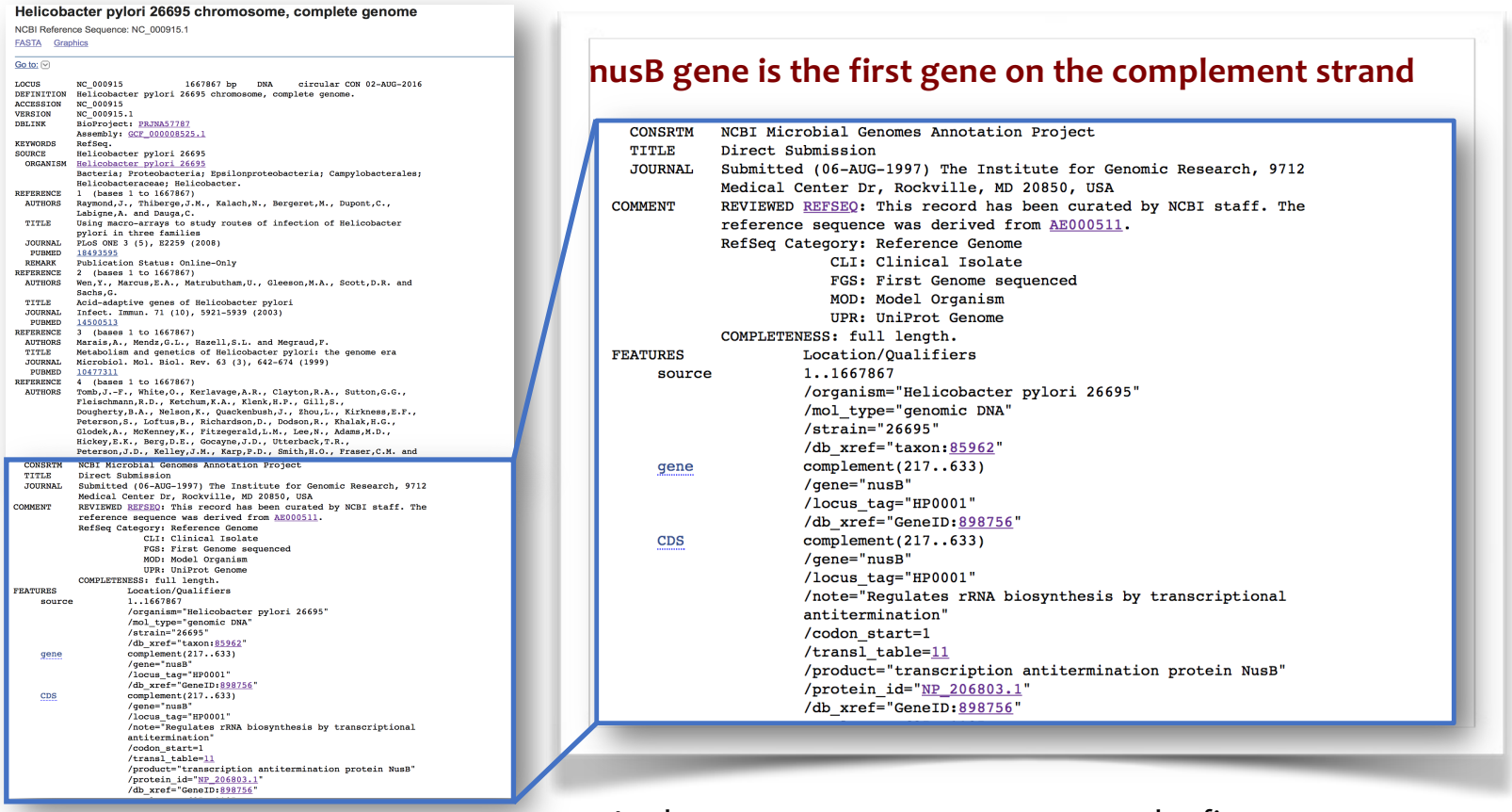
- For example, the assembled sequence can start at S1, S2, S3 or S4.

So re-ordering the start point of the genome sequence is needed!

How to do it in a proper way?

- First gene selection etc.

The First Gene in HP26695 Genome

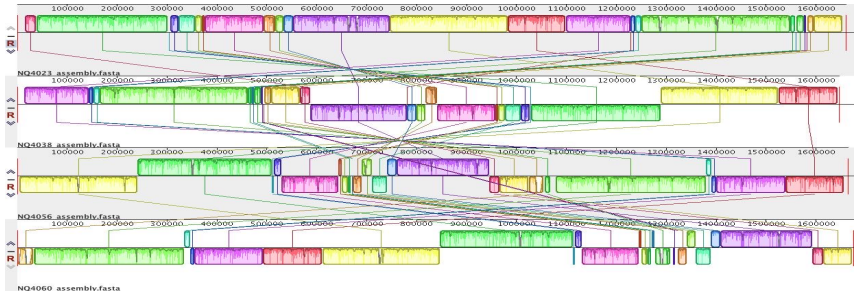


In theory, you can use any gene as the first gene.

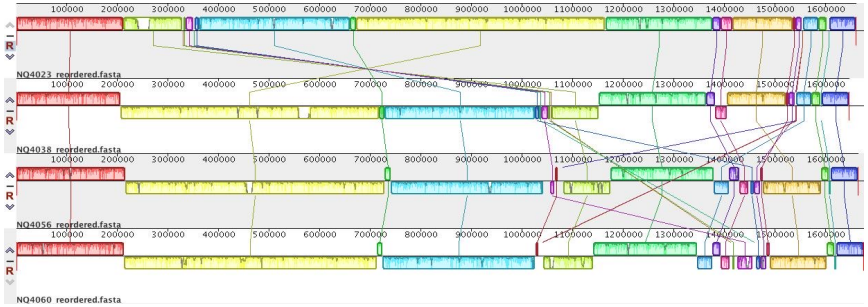
Re-ordering the Genome

1. Find the nusB gene in the assembled genome → 2. Set it as the first gene in genome → 3. Make the cut and join two pieces of sequences together

Before re-order



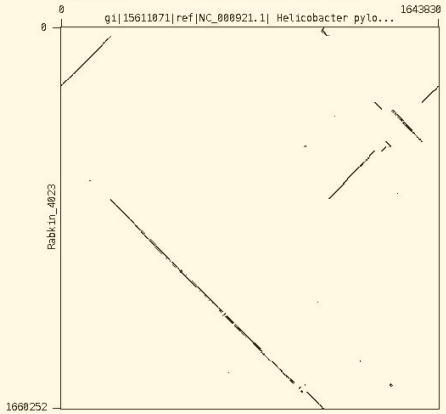
After re-order



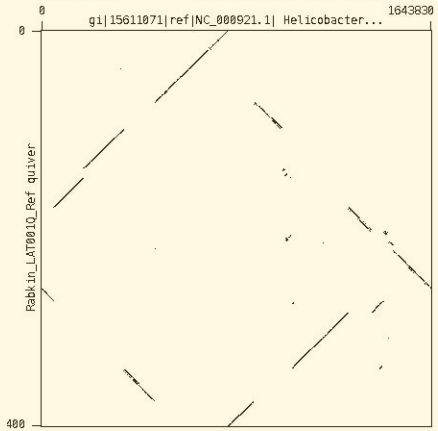
Dr. Kasia Thorell
Karolinska Institute

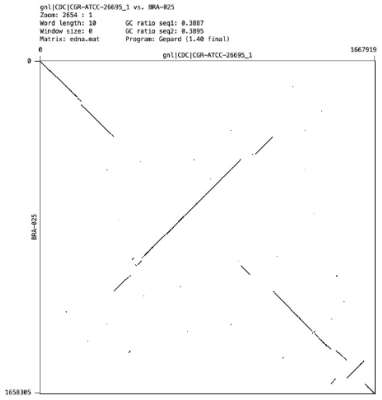
DOT plot

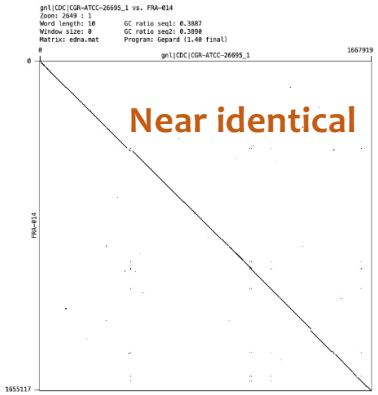
HpyJ99 vs. COL4023



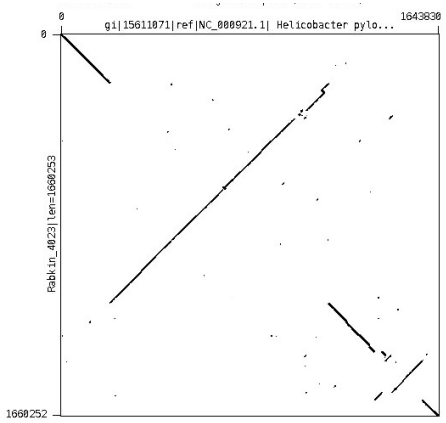
HpyJ99 vs. LAT001



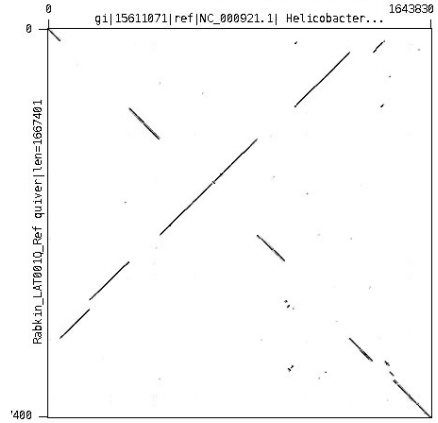




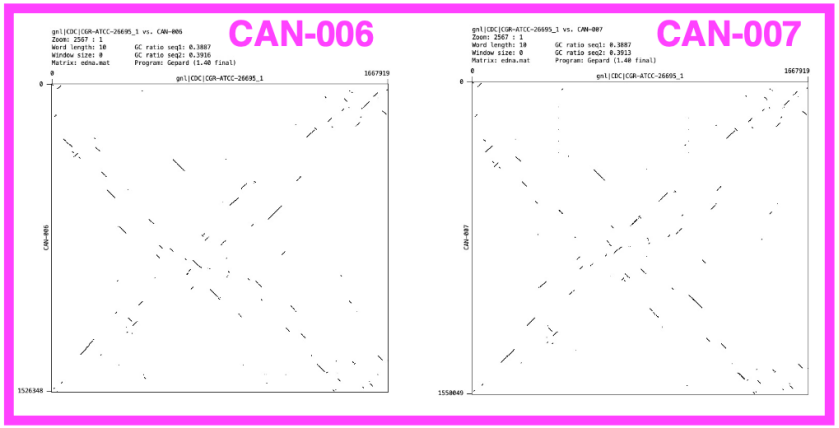
HpyJ99 vs. COL4023 fixed



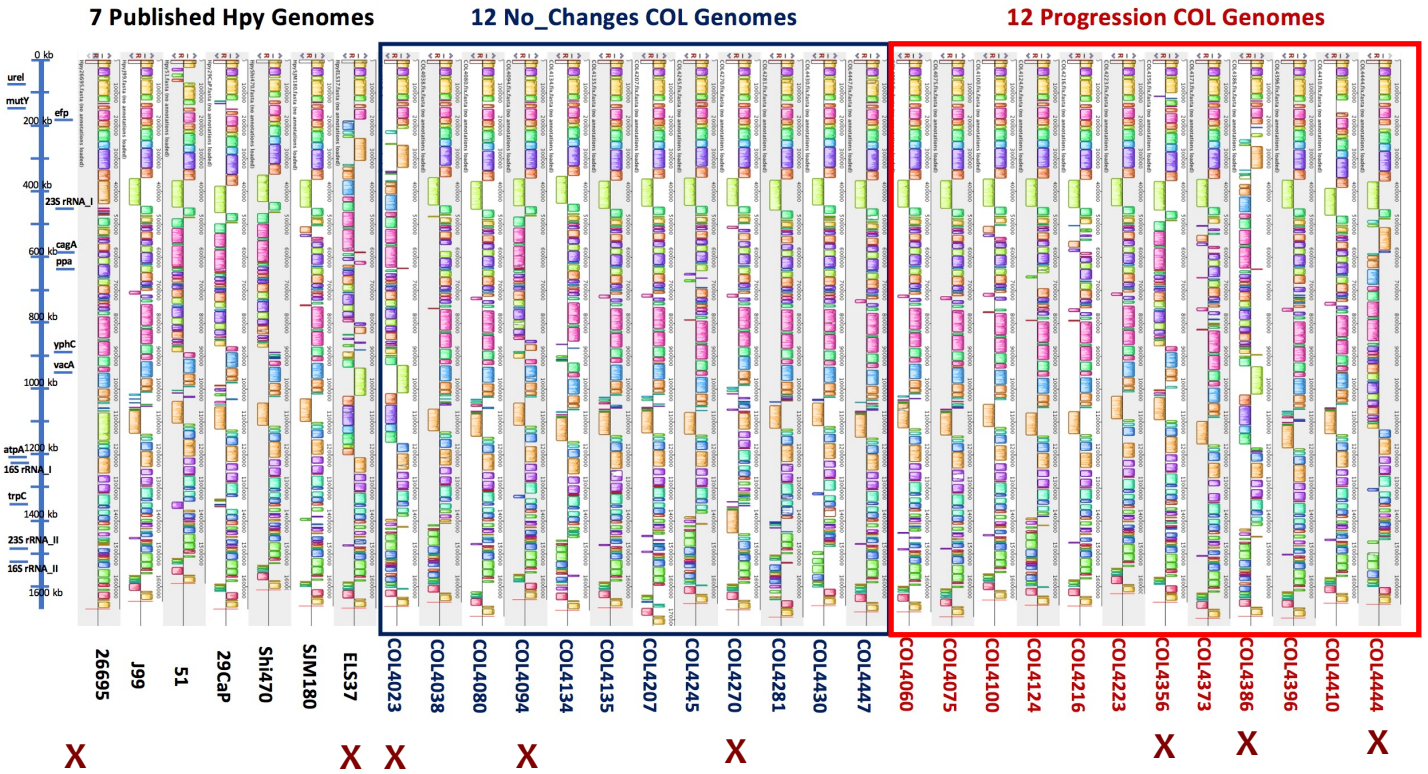
HpyJ99 vs. LAT001 fixed



Outliers



7 Published and 24 COL Genomes Alignment



Mauve genome alignment

Other Major Issues and Challenges

- In 2017, most groups used Illumina technology to do *de novo* assembly or align reads to HP26695.

  Quite often, they got incomplete assemblies.

- How do we check the sequencing quality?

- How do we evaluate these *de novo* assemblies?

- How do we find the unintentional duplicates in the dataset?

- How to evaluate plasmids?

Prof. Ichizo Kobayashi, University of Tokyo

Dr. Richard Roberts, New England Biolabs

Internal investigators

Wen, Kedest, Kristie
Josh Cherry(NCBI), John Dekker(NIAID)

Other Major Issues and Challenges

- In 2017, most groups used Illumina technology to do *de novo* assembly or align reads to HP26695.

  Quite often, they got incomplete assemblies.

  - Comparison of the Illumina vs PacBio results
- How do we check the sequencing quality?
- How do we evaluate these *de novo* assemblies?
- How do we find the unintentional duplicates in the dataset?
- How to evaluate plasmids?

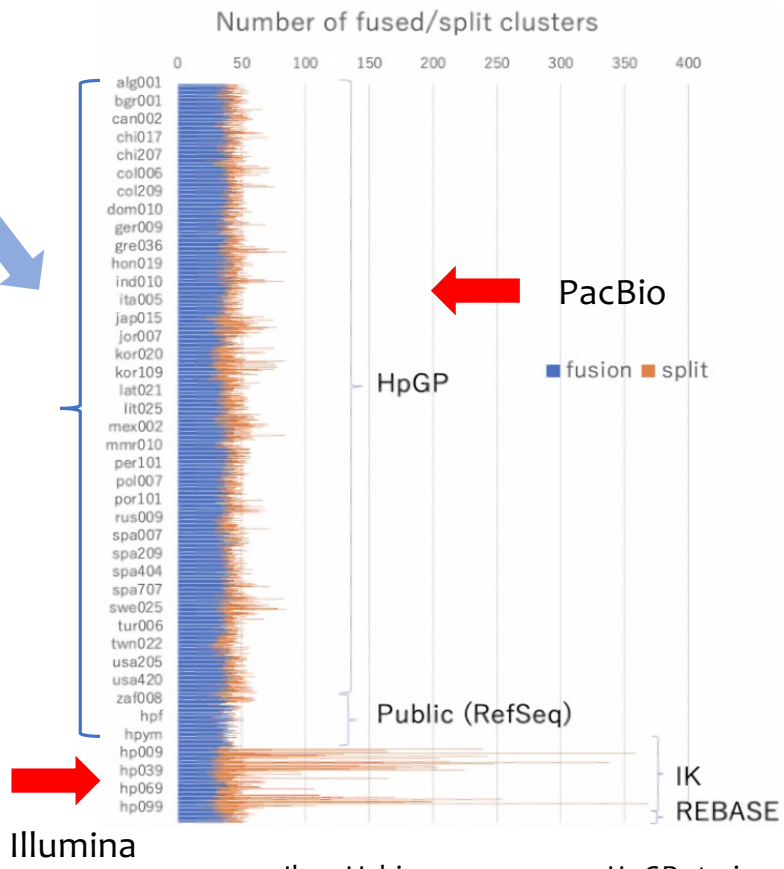
Prof. Ichizo Kobayashi, University of Tokyo

Dr. Richard Roberts, New England Biolabs

Internal investigators

Wen, Kristie

HpGP, RefSeq, IK and REBASE



Ikuo Uchiyama

~750 HpGP strains

Other Major Issues and Challenges

- In 2017, most groups used Illumina technology to do *de novo* assembly or align reads to HP26695.

  Quite often, they got incomplete assemblies.

  - Comparison of the Illumina vs. PacBio results

- How do we check the sequencing quality?

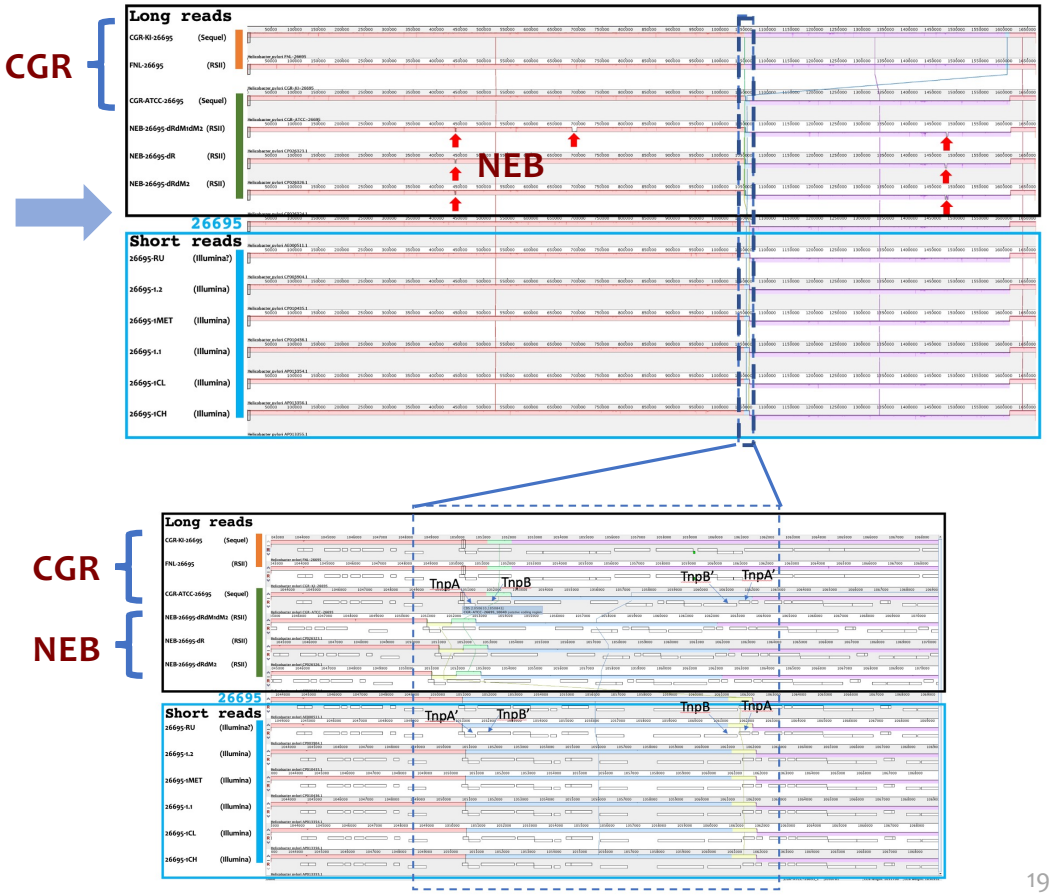
  - Resequencing the strains with known sequences

- How do we evaluate these *de novo* assemblies?

- How do we find the unintentional duplicates in the dataset?

- How to evaluate plasmids?

HP26695: Long reads vs. Short reads



Prof. Ichizo Kobayashi, University of Tokyo

Dr. Richard Roberts, New England Biolabs

Internal investigators

Wen, Kristie

Other Major Issues and Challenges

- In 2017, most groups used Illumina technology to do *de novo* assembly or align reads to HP26695.

  Quite often, they got incomplete assemblies.

  - Comparison of the Illumina vs. PacBio results
- How do we check the sequencing quality?
  - Resequencing the strains with known sequences
- How do we evaluate these *de novo* assemblies?
  - BUSCO score & number of pseudo genes
- How do we find the unintentional duplicates in the dataset?
- How to evaluate plasmids?

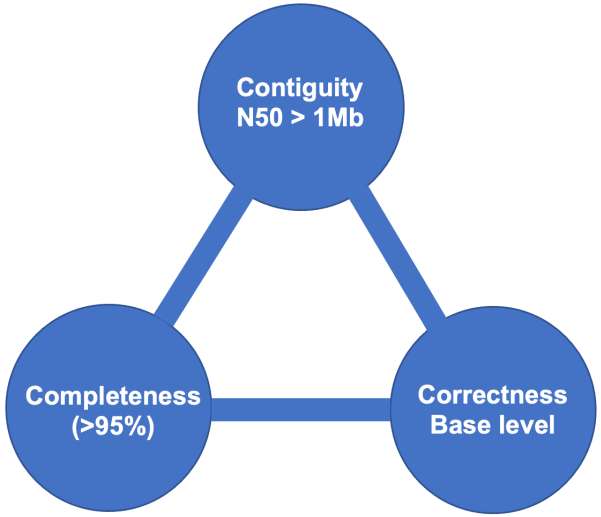
Prof. Ichizo Kobayashi, University of Tokyo

Dr. Richard Roberts, New England Biolabs

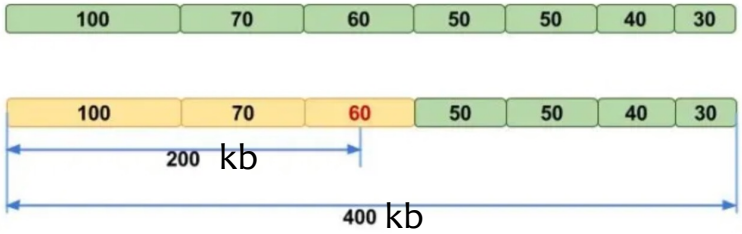
Internal investigators

Wen, Kedest, Kristie

Assessing the Quality of Genome Assemblies with the 3C's



N50 explained



https://www.pacb.com/blog/beyond-contiguity
https://www.moleculareco logist.com/2017/03/29/whats-n50/

Contiguity is often measured as contig N50, which is the length cutoff for the longest contigs that contain 50% of the total genome length. **In this era of long-read genome assemblies, a contig N50 over 1 Mb is generally considered good.**

Completeness is often measured using BUSCO (Benchmarking Universal Single-Copy Orthologs) scores, which look for the presence or absence of highly conserved genes in an assembly. The aim is to have the highest percentage of genes identified in your assembly, with **a BUSCO complete score above 95% considered good.**

Correctness, the third and final C, **is more challenging to measure. Correctness can be defined as the accuracy of each base pair in the assembly** and is most often measured as concordance of an assembly to a gold standard reference. Of course, when sequencing a novel species there may not be a reference against which to measure. Furthermore, **concordance is only a good measure for accuracy when the gold-standard itself is very high quality and when there is little biological divergence between the reference sample and assembly sample.**

Completeness of Genome Assemblies

Genome analysis

BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs

Felipe A. Simão†, Robert M. Waterhouse†, Panagiotis Ioannidis, Evgenia V. Kriventseva and Evgeny M. Zdobnov*

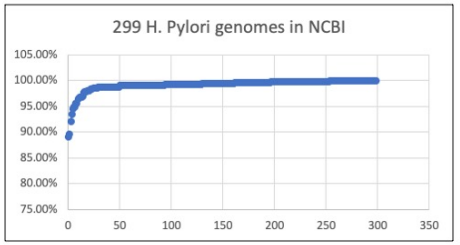
Department of Genetic Medicine and Development, University of Geneva Medical School and Swiss Institute of Bioinformatics, rue Michel-Servet 1, 1211 Geneva, Switzerland

We propose intuitive metrics to describe genome, gene set or transcriptome completeness in BUSCO notation - C:complete [D:duplicated], F:fragmented, M:missing, n:number of genes used (Fig. 1). The recovered genes are classified as 'complete' when their lengths are within two standard deviations of the BUSCO group mean length (i.e. within ~95% expectation, Supplementary Fig. S1). 'Complete' genes found with more than one copy are classified as 'duplicated'. These should be rare, as BUSCOs are evolving under single-copy control (Waterhouse *et al.*, 2011), and the recovery of many duplicates may therefore indicate erroneous assembly of haplotypes. Genes only partially recovered are classified as 'fragmented', and genes not recovered are classified as 'missing'. Finally, the 'number of genes used' indicates the resolution and hence is informative of the confidence of these assessments.

Complete BUSCOs (C)
Complete and duplicated BUSCOs (D)
Complete and single-copy BUSCOs (S)
Fragmented BUSCOs (F)
Missing BUSCOs (M)
Total BUSCO groups searched (n)

AccessionID	C	S	D	F	M	n
CP032043	89.00%	89.00%	0.00%	7.50%	3.50%	628
AP014711	89.60%	89.60%	0.00%	6.70%	3.70%	628
CP032039	92.00%	91.70%	0.30%	5.60%	2.40%	628
CP032027	93.40%	93.20%	0.20%	3.80%	2.80%	628
CP003419	94.40%	94.40%	0.00%	3.70%	1.90%	628
CP023265	94.70%	94.70%	0.00%	3.50%	1.80%	628
CP032046	94.70%	93.90%	0.80%	4.10%	1.20%	628
AP014712	95.40%	95.40%	0.00%	3.50%	1.10%	628
CP006691	95.50%	95.50%	0.00%	3.30%	1.20%	628
CP032041	96.40%	96.20%	0.20%	1.90%	1.70%	628
CP024017	96.50%	96.50%	0.00%	1.80%	1.70%	628
AP014710	96.70%	96.70%	0.00%	2.10%	1.20%	628
CP023266	96.70%	96.70%	0.00%	2.10%	1.20%	628
CP024015	96.70%	96.70%	0.00%	1.90%	1.40%	628
CP032038	96.90%	96.70%	0.20%	2.40%	0.70%	628

	C	S	D	F	M
min	89.00%	89.00%	0.00%	0.00%	0.10%
median	99.40%	99.40%	0.00%	0.20%	0.30%
max	99.90%	99.80%	3.20%	7.50%	3.70%



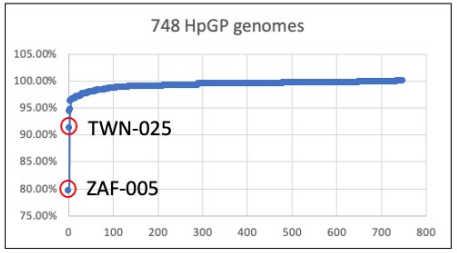
GenBank

Lab resequencing
Or *in silico* polishing?

All passed Contiguity assessment (N50 > 1Mb)

sampleID	C	S	D	F	M	n
ZAF-005	79.50%	79.50%	0.00%	11.00%	9.50%	628
TWN-025	91.20%	90.90%	0.30%	6.20%	2.60%	628
COL-007	94.40%	94.40%	0.00%	2.40%	3.20%	628
COL-304	94.60%	94.60%	0.00%	3.20%	2.20%	628
KOR-035	96.20%	96.20%	0.00%	2.20%	1.60%	628
IND-006	96.30%	96.30%	0.00%	1.40%	2.30%	628
KOR-002	96.30%	96.30%	0.00%	2.70%	1.00%	628
JAP-104	96.50%	96.50%	0.00%	2.10%	1.40%	628
KOR-012	96.50%	96.50%	0.00%	2.70%	0.80%	628
KOR-110	96.50%	96.50%	0.00%	2.40%	1.10%	628
SWE-024	96.50%	96.50%	0.00%	2.10%	1.40%	628
SWE-026	96.50%	96.50%	0.00%	2.40%	1.10%	628
JAP-105	96.70%	96.70%	0.00%	1.80%	1.50%	628
KOR-041	96.70%	96.70%	0.00%	2.10%	1.20%	628
KOR-048	96.70%	96.70%	0.00%	2.10%	1.20%	628

	C	S	D	F	M
min	79.50%	79.50%	0.00%	0.00%	0.00%
median	99.50%	99.40%	0.00%	0.20%	0.30%
max	100.00%	100.00%	2.10%	11.00%	9.50%



HpGP

https://busco-data.ezlab.org/v5/data/lineages/campylobacterales_odb10.2020-03-06.tar.gz

Identified 4 genomes need to be improved.

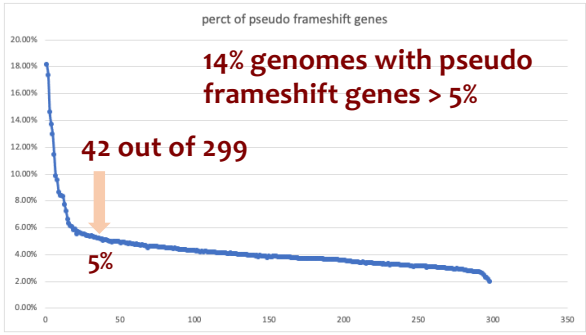
Batch3: 748 strains

Pseudogenes

299 NCBI GenBank *H. pylori* Genomes (Complete or Chromosome only) (2021-08)

strain	LOCUS	ACCESSION	length	Gene_all	CDS_all	Gene_coding	CDS	Gene_RNA	rRNA	lete_rRNAs	tRNAs	ncRNAs	Pseudo_all	Pseudo_CDS	_ambiguous	udo_framshifted	incomplete	seudo_stop	udo_others	perct	pert2
GCF_001549855.1_	NZ_AP014711	NZ_AP014711	1562125	1545	1500	1204	1204	45	2,2,2(5S	2,2,2(5S,16	36	3	296	296	0	272	18	20	13	18.13%	19.73%
GCF_008033035.1_	NZ_CP032043	NZ_CP032043	1632682	1594	1549	1245	1245	45	2,2,2(5S	2,2,2(5S,16	36	3	304	304	0	269	32	18	15	17.37%	19.63%
GCF_008032935.1_	NZ_CP032039	NZ_CP032039	1645512	1620	1575	1314	1314	45	2,2,2(5S	2,2,2(5S,16	36	3	261	261	0	230	28	27	24	14.60%	16.57%
GCF_000262655.1_	NC_017926	NC_017926	1634138	1647	1602	1345	1345	45	2,2,2(5S	2,2,2(5S,16	36	3	257	257	0	219	41	28	29	13.67%	16.04%
GCF_008032955.1_	NZ_CP032046	NZ_CP032046	1632652	1603	1558	1325	1325	45	2,2,2(5S	2,2,2(5S,16	36	3	233	233	0	202	19	33	20	12.97%	14.96%
GCF_002952235.1_	NZ_CP023265	NZ_CP023265	1674350	1587	1542	1345	1345	45	2,2,2(5S	2,2,2(5S,16	36	3	197	197	0	176	21	15	14	11.41%	12.78%
GCF_001549875.1_	NZ_AP014712	NZ_AP014712	1629114	1574	1531	1353	1353	43	1,2,1(5S	1,2,1(5S,16	36	3	178	178	0	151	22	17	12	9.86%	11.63%
GCF_000590775.1_	NC_022130	NC_022130	1622903	1568	1523	1342	1342	45	2,2,2(5S	2,2,2(5S,16	36	3	181	181	0	145	34	22	19	9.52%	11.88%
GCF_002952335.1_	NZ_CP024015	NZ_CP024015	1674120	1639	1594	1430	1430	45	2,2,2(5S	2,2,2(5S,16	36	3	164	164	0	138	14	24	12	8.66%	10.29%
GCF_001549715.1_	NZ_AP014710	NZ_AP014710	1629815	1566	1521	1369	1369	45	2,2,2(5S	2,2,2(5S,16	36	3	152	152	0	128	14	20	9	8.42%	9.99%
GCF_002952375.1_	NZ_CP024017	NZ_CP024017	1674163	1641	1596	1436	1436	45	2,2,2(5S	2,2,2(5S,16	36	3	160	160	0	134	17	23	14	8.40%	10.03%
GCF_008032615.1_	NZ_CP032038	NZ_CP032038	1661266	1632	1587	1422	1422	45	2,2,2(5S	2,2,2(5S,16	36	3	165	165	0	132	27	20	13	8.32%	10.40%
GCF_002952255.1_	NZ_CP023266	NZ_CP023266	1674214	1576	1531	1396	1396	45	2,2,2(5S	2,2,2(5S,16	36	3	135	135	0	118	14	14	10	7.71%	8.82%
GCF_008032735.1_	NZ_CP032036	NZ_CP032036	1624459	1528	1483	1360	1360	45	2,2,2(5S	2,2,2(5S,16	36	3	123	123	0	107	11	12	4	7.22%	8.29%
GCF_008032995.1_	NZ_CP032041	NZ_CP032041	1697806	1630	1585	1431	1431	45	2,2,2(5S	2,2,2(5S,16	36	3	154	154	0	105	41	20	12	6.62%	9.72%
GCF_003640585.1_	NZ_CP031558	NZ_CP031558	1581994	1525	1480	1370	1370	45	2,2,2(5S	2,2,2(5S,16	36	3	110	110	0	93	16	14	12	6.28%	7.43%
GCF_002356595.1_	NZ_CP023448	NZ_CP023448	1578351	1523	1478	1367	1367	45	2,2,2(5S	2,2,2(5S,16	36	3	111	111	0	91	12	17	9	6.16%	7.51%
GCF_002952355.1_	NZ_CP024016	NZ_CP024016	1674080	1583	1538	1431	1431	45	2,2,2(5S	2,2,2(5S,16	36	3	107	107	0	93	11	14	10	6.05%	6.96%
GCF_001433495.1_	NZ_CP012907	NZ_CP012907	1667159	1590	1544	1420	1420	46	3,2,2(5S	3,2,2(5S,16	36	3	124	124	0	90	20	24	10	5.83%	8.03%
GCF_000021465.1_	NC_011498	NC_011498	1673813	1597	1552	1441	1441	45	2,2,2(5S	2,2,2(5S,16	36	3	111	111	0	91	21	14	14	5.86%	7.15%
GCF_003711165.1_	NZ_CP025748	NZ_CP025748	1667396	1535	1489	1301	1301	46	3,2,2(5S	3,2,2(5S,16	36	3	188	188	72	82	48	31	37	5.51%	12.63%
GCF_002357755.1_	NZ_AP017355	NZ_AP017355	1566172	1512	1467	1363	1363	45	2,2,2(5S	2,2,2(5S,16	36	3	104	104	9	84	12	22	22	5.73%	7.09%
GCF_002357475.1_	NZ_AP017331	NZ_AP017331	1607914	1522	1475	1362	1362	47	3,2,3(5S	3,2,3(5S,16	36	3	113	113	14	83	19	17	19	5.63%	7.66%
GCF_008032475.1_	NZ_CP032037	NZ_CP032037	1667329	1581	1533	1431	1431	48	2,2,2(5S	2,2,2(5S,16	39	3	102	102	0	86	13	8	5	5.61%	6.65%
GCF_002357335.1_	NZ_AP017339	NZ_AP017339	1616737	1535	1490	1385	1385	45	2,2,2(5S	2,2,2(5S,16	36	3	105	105	13	82	13	17	19	5.50%	7.05%
GCF_002357575.1_	NZ_AP017336	NZ_AP017336	1566831	1514	1469	1371	1371	45	2,2,2(5S	2,2,2(5S,16	36	3	98	98	5	81	12	18	17	5.51%	6.67%
GCF_000148665.1_	NC_017357	NC_017357	1549666	1492	1450	1339	1339	42	1,1,1(5S	1,1,1(5S,16	36	3	111	111	0	79	26	17	9	5.45%	7.66%
GCF_000192335.1_	NC_017381	NC_017381	1562832	1502	1460	1354	1354	42	1,1,1(5S	1,1,1(5S,16	36	3	106	106	0	79	20	17	8	5.41%	7.26%
GCF_002357635.1_	NZ_AP017345	NZ_AP017345	1623153	1574	1527	1413	1413	47	3,2,3(5S	3,2,3(5S,16	36	3	114	114	11	82	18	25	22	5.37%	7.47%
GCF_002357415.1_	NZ_AP017352	NZ_AP017352	1644960	1582	1535	1415	1415	47	3,2,3(5S	3,2,3(5S,16	36	3	120	120	11	82	19	29	20	5.34%	7.82%
GCF_000192315.1_	NC_017374	NC_017374	1548238	1495	1453	1347	1347	42	1,1,1(5S	1,1,1(5S,16	36	3	106	106	0	78	20	16	6	5.37%	7.30%
GCF_008026175.1_	NZ_CP032020	NZ_CP032020	1683530	1618	1573	1471	1471	45	2,2,2(5S	2,2,2(5S,16	36	3	102	102	0	84	17	9	8	5.34%	6.48%
GCF_000213135.1_	NC_017375	NC_017375	1617426	1546	1501	1397	1397	45	2,2,2(5S	2,2,2(5S,16	36	3	104	104	0	79	13	23	10	5.26%	6.93%
GCF_900638505.1_	NZ_LR134519	NZ_LR134519	1632224	1545	1500	1408	1408	45	2,2,2(5S	2,2,2(5S,16	36	3	92	92	0	79	8	10	5	5.27%	6.13%
GCF_008026615.1_	NZ_CP032033	NZ_CP032033	1645502	1559	1514	1416	1416	45	2,2,2(5S	2,2,2(5S,16	36	3	98	98	0	79	10	15	6	5.22%	6.47%
GCF_008033055.1_	NZ_CP032048	NZ_CP032048	1672280	1563	1517	1430	1430	46	2,2,2(5S	2,2,2(5S,16	37	3	87	87	0	79	5	8	4	5.21%	5.74%
GCF_002906515.1_	NZ_CP026323	NZ_CP026323	1666879	1596	1551	1457	1457	45	2,2,2(5S	2,2,2(5S,16	36	3	94	94	0	80	13	9	8	5.16%	6.06%
GCF_008032975.1_	NZ_CP032040	NZ_CP032040	1613306	1528	1483	1397	1397	45	2,2,2(5S	2,2,2(5S,16	36	3	86	86	0	76	7	10	6	5.12%	5.80%
GCF_002357715.1_	NZ_AP017351	NZ_AP017351	1627132	1554	1509	1402	1402	45	2,2,2(5S	2,2,2(5S,16	36	3	107	107	8	76	16	19	12	5.04%	7.09%
GCF_000008525.1_	NC_000915	NC_00915 NZ_	1667867	1584	1539	1442	1442	45	2,2,2(5S	2,2,2(5S,16	36	3	97	97	0	78	13	14	8	5.07%	6.30%
GCF_002357595.1_	NZ_AP017337	NZ_AP017337	1558697	1507	1462	1373	1373	45	2,2,2(5S	2,2,2(5S,16	36	3	89	89	5	74	17	11	16	5.06%	6.09%
GCF_006338285.1_	NZ_CP032910	NZ_CP032910	1599658	1519	1474	1382	1382	45	2,2,2(5S	2,2,2(5S,16	36	3	92	92	0	74	16	15	13	5.02%	6.24%
GCF_001653475.1_	NZ_CP011487	NZ_CP011487	1531450	1484	1442	1346	1346	42	1,1,1(5S	1,1,1(5S,16	36	3	96	96	0	72	22	22	19	4.99%	6.66%
GCF_900119995.1_	NZ_LT635458	NZ_LT635458	1667804	1588	1543	1444	1444	45	2,2,2(5S	2,2,2(5S,16	36	3	99	99	0	77	18	14	7	4.99%	6.42%
GCF_002357375.1_	NZ_AP017346	NZ_AP017346	1588584	1527	1480	1378	1378	47	3,2,3(5S	3,2,3(5S,16	36	3	102	102	12	73	17	18	16	4.93%	6.89%
GCF_002357315.1_	NZ_AP017360	NZ_AP017360	1606460	1543	1496	1397	1397	47	3,2,3(5S	3,2,3(5S,16	36	3	99	99	8	74	13	24	20	4.95%	6.62%
GCF_002357855.1_	NZ_AP017362	NZ_AP017362	1577123	1516	1471	1379	1379	45	2,2,2(5S	2,2,2(5S,16	36	3	92	92	12	73	12	10	15	4.96%	6.25%
GCF_002222575.1_	NZ_CP022409	NZ_CP022409	1576133	1515	1469	1381	1381	46	2,2,2(5S	2,2,2(5S,16	37	3	88	88	0	73	11	11	7	4.97%	5.99%
GCF_001433515.1_	NZ_CP012905	NZ_CP012905	1624441	1539	1494	1405	1405	45	2,2,2(5S	2,2,2(5S,16	36	3	89	89	0	74	15	4	4	4.95%	5.96%

Percent of pseudo frameshift genes



Pseudogenes

748 HpGP Genomes (Batch3)

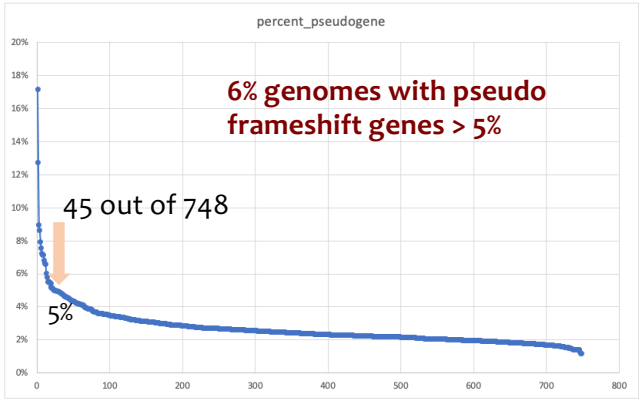
sampleID	prokka_CDS	prokka_CDS_	percent_pseudogene	tRNA	16SrRNA	23SrRNA	length	NCBI_CDS	NCBI_frames	perct
ZAF-005	1926	331	17%	36	2	1	1650827	NA	NA	
POR-114	1735	221	13%	36	2	2	1635540	NA	NA	
TWN-025	1670	150	9%	37	2	2	1625574	1560	249	16%
COG-011	1630	141	9%	36	2	2	1615291	NA	NA	
SWE-001	1648	131	8%	36	2	2	1606996	1553	165	11%
KOR-001	1648	125	8%	36	2	2	1639021	1559	164	11%
SWE-016	1645	119	7%	36	2	2	1617261	1557	183	12%
COG-005	1608	115	7%	36	2	2	1615710	NA	NA	
COL-304	1641	117	7%	36	2	2	1654360	NA	NA	
COL-007	1655	113	7%	36	2	2	1653789	NA	NA	
PER-016	1661	110	7%	36	2	2	1689329	NA	NA	
MAL-006	1649	109	7%	36	2	2	1664187	1596	152	10%
KOR-044	1592	96	6%	36	2	2	1607983	1541	152	10%
CHI-010	1574	91	6%	36	2	2	1602017	NA	NA	
KOR-004	1554	86	6%	36	2	2	1578546	1513	123	8%
SWE-031	1616	89	6%	36	2	2	1643010	1560	136	9%
KOR-005	1600	88	6%	36	2	2	1634969	1555	122	8%
KOR-002	1568	85	5%	36	2	2	1597612	1539	122	8%
KOR-045	1591	86	5%	36	2	2	1609731	NA	NA	
IND-006	1622	84	5%	36	2	2	1670468	NA	NA	
KOR-035	1554	80	5%	36	2	2	1581821	NA	NA	
JAP-002	1543	79	5%	36	2	2	1562309	NA	NA	
JAP-109	1550	78	5%	36	2	2	1585338	NA	NA	
KOR-048	1533	77	5%	36	2	2	1574538	1504	103	7%
MEX-028	1535	77	5%	36	2	2	1562280	1481	116	8%
KOR-041	1578	79	5%	36	2	2	1605825	1548	133	9%
SWT-007	1627	81	5%	36	2	2	1663034	1584	134	8%
KOR-110	1577	78	5%	36	2	2	1623260	NA	NA	
JAP-104	1578	78	5%	36	2	2	1629601	NA	NA	
KOR-010	1549	76	5%	36	2	2	1582866	1513	123	8%
SGP-018	1593	78	5%	36	2	2	1634713	NA	NA	
SWE-024	1577	77	5%	36	2	2	1625484	1560	122	8%
KOR-033	1577	76	5%	36	2	2	1613357	1548	133	9%
COG-004	1560	75	5%	36	2	2	1615896	NA	NA	
JAP-010	1536	73	5%	36	2	2	1574397	1516	118	8%
JAP-105	1518	72	5%	36	2	2	1555420	NA	NA	
MAL-023	1539	72	5%	36	2	2	1582641	1518	86	6%
SWE-026	1604	75	5%	36	2	2	1654901	1567	148	9%
SWE-004	1532	71	5%	36	2	2	1600123	1547	123	
GRE-041	1561	72	5%	36	2	2	1626901	NA	NA	
SWE-022	1592	73	5%	36	2	2	1644054	1577	131	
MAL-020	1532	70	5%	36	2	2	1583392	1513	113	7%
PRI-001	1515	69	5%	36	2	2	1577958	NA	NA	
SPA-611	1606	73	5%	36	2	2	1666518	NA	NA	
JAP-111	1533	69	5%	36	2	2	1563824	NA	NA	
GRE-044	1613	72	4%	36	2	2	1688552			

17

28

45 genomes >= 5%; 17 genomes > 5%

Based on Prokka annotation



tRNA: 36
16SrRNA: 2
23SrRNA: 2
% Pseudo gene: <= 5%

These 17 genomes need to be updated.

Other Major Issues and Challenges

- In 2017, most groups used Illumina technology to do *de novo* assembly or align reads to HP26695.

  Quite often, they got incomplete assemblies.

  - Comparison of the Illumina vs. PacBio results
- How do we check the sequencing quality?
  - Resequencing the strains with known sequences
- How do we evaluate these *de novo* assemblies?
  - BUSCO score
- How do we find the unintentional duplicates in the dataset?
  - Use Ichizo's 20 family strains set
- How to evaluate plasmids?

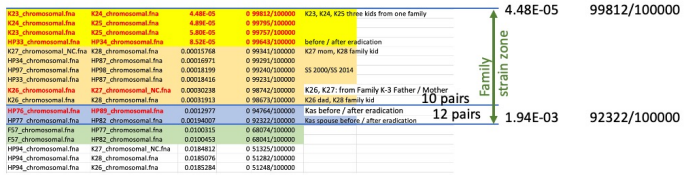
Prof. Ichizo Kobayashi, University of Tokyo

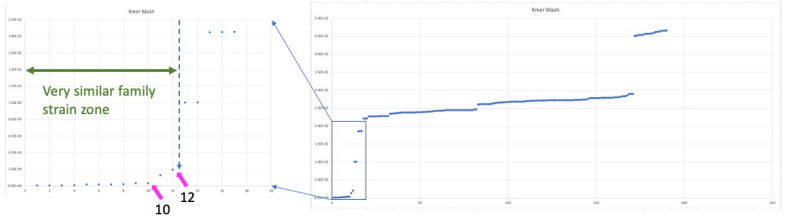
Dr. Richard Roberts, New England Biolabs

Internal investigators

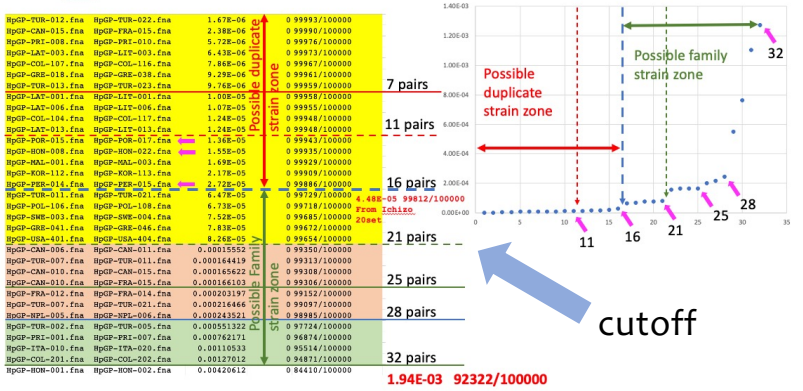
Wen, Kedest, Kristie

Ichizo's 20 family strains mash kmer





HpGP strains mash kmer



Other Major Issues and Challenges

- In 2017, most groups used Illumina technology to do de novo assembly or align reads to HP26695.

  Quite often, they got incomplete assemblies.

  - Comparison of the Illumina vs. PacBio results
- How do we check the sequencing quality?
  - Resequencing the strains with known sequences
- How do we evaluate these de novo assemblies?
  - BUSCO score
- How do we find the unintentional duplicates in the dataset?
  - Use Ichizo's 20 family strains set
- How to evaluate plasmids?
  - Exonuclease V digestion screening

Prof. Ichizo Kobayashi, University of Tokyo

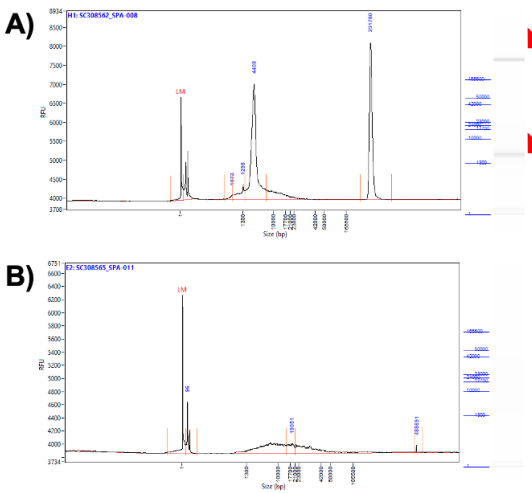
Dr. Richard Roberts, New England Biolabs

Internal investigators

Wen, Kedest, Kristie
Josh Cherry(NCBI), John Dekker(NIAID)

Exonuclease V digestion screening

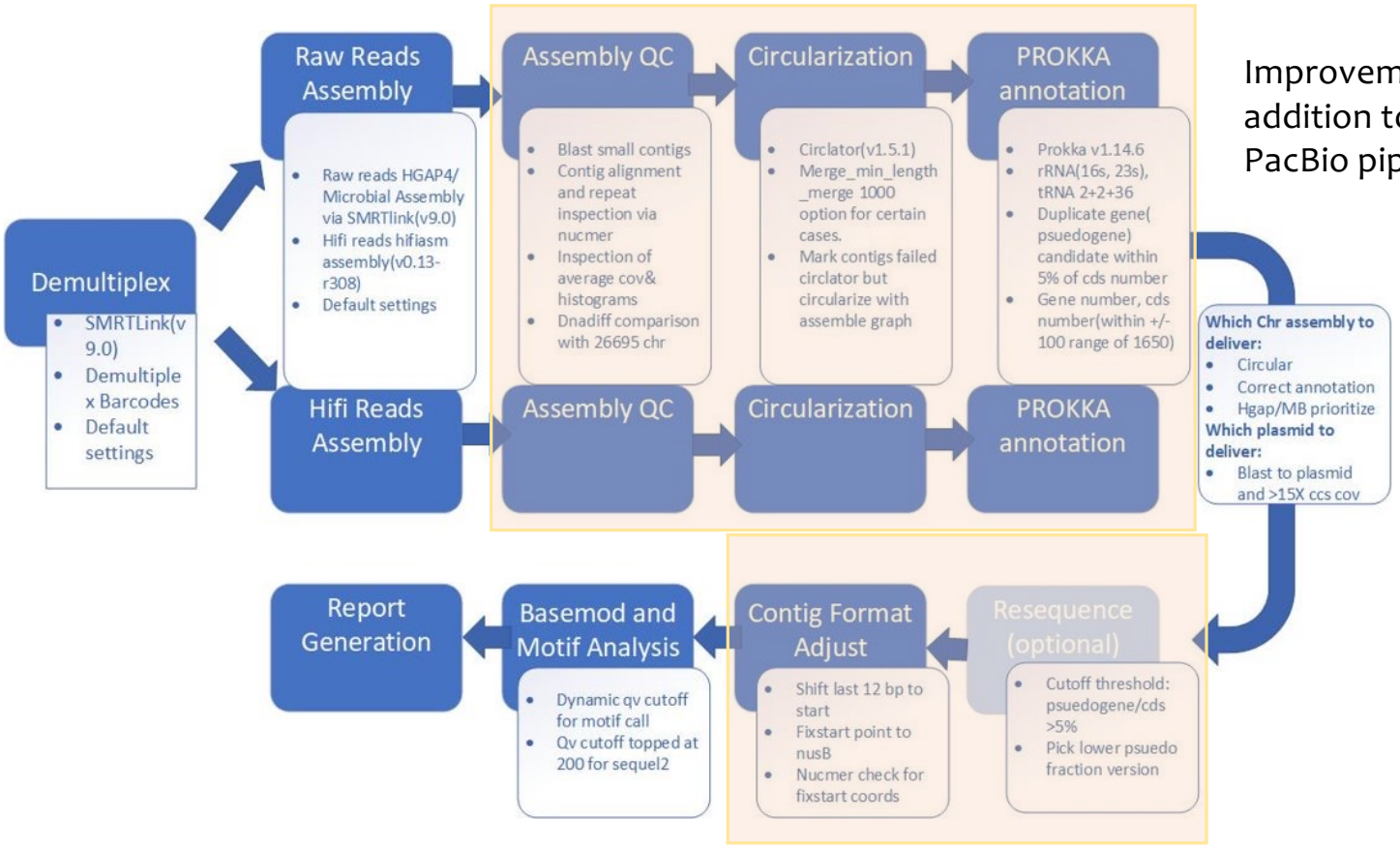
Status	Sample Number
Exonuclease screening performed	623
Flagged for follow-up assembly	129
Plasmid contig identified in follow-up assembly	52



Initially, no plasmid detected.
After screening, plasmids detected

Initially, no plasmid detected.
After screening, no plasmids detected

Improved *H. pylori* Genome Assembly Pipeline



Improvement steps in addition to the original PacBio pipeline

Wen and Kristie

HpGP Data Analyses

Primary Analyses

- Population structure and ancestry (Thorell, Muñoz-Ramirez *et al.*) Landmark paper published.
- Prophages (Vale *et al.*) Under review.
- Methylation profiles (Roberts *et al.*) in preparation.
- Plasmids (Torres R. *et al.*) ongoing.
- Antibiotic resistance to clarithromycin and levofloxacin (Chiner-Oms *et al.*) ongoing.
- Genome- and epigenome wide association analyses of gastric cancer and advanced intestinal metaplasia (Yahara *et al.* → Wang) ongoing.

HpGP data (1012 genomes) was finalized in Arpil 2022.

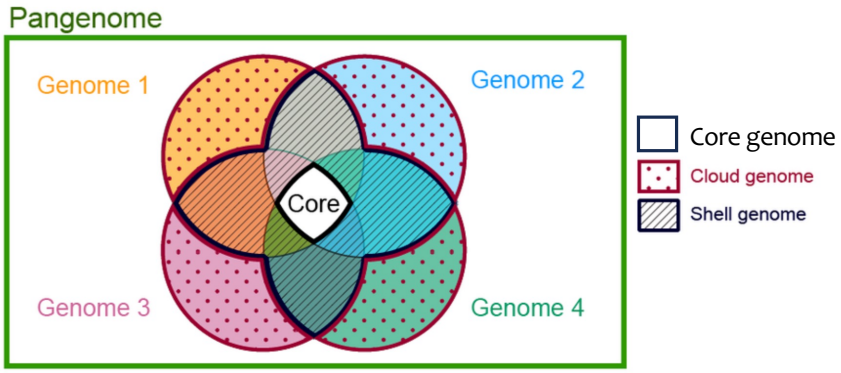
Secondary Analyses

- Rearrangements
- Integrative conjugative elements
- Gene-centric
  - vacA, babA, babB, babC, hopQ, cagY, 16S
- Metabolic pathways
- Mutational signature
- Adaptive differentiation
- Non-coding RNAs

Focus of the *H. pylori* landmark paper

Pan and Core Genome Analyses

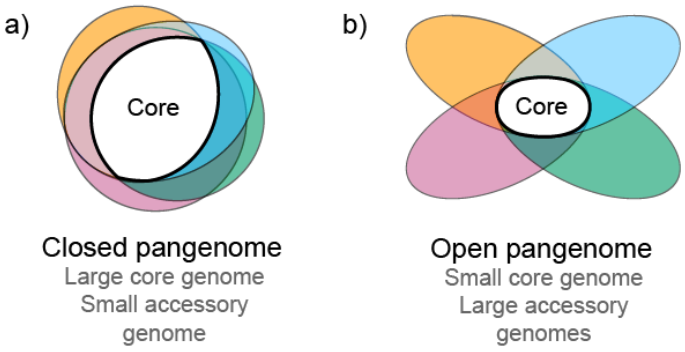
- A pan-genome (pangenome or supragenome) is the entire set of genes from all strains within a clade.

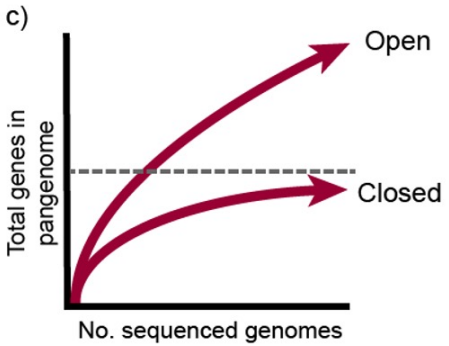


The $\alpha = 0.879 \pm 0.035 < 1$ (> 30 genomes) value of the Heaps' law indicates that **the pan-genome of *H. pylori* is "open"** i.e., the size of the pan-genome tends to diverge when N increases, as concluded in a previous analysis using seven *H. pylori* genomes.

Uchiyama et al. PLoS ONE 11(8): e0159419, (2016)
Tettelin et al. Curr. Opin. Microbiol. 11, 472-477 (2008)
Fischer W et al. NAR, 38, 6089-6101, (2010)

- Open and closed pangenome





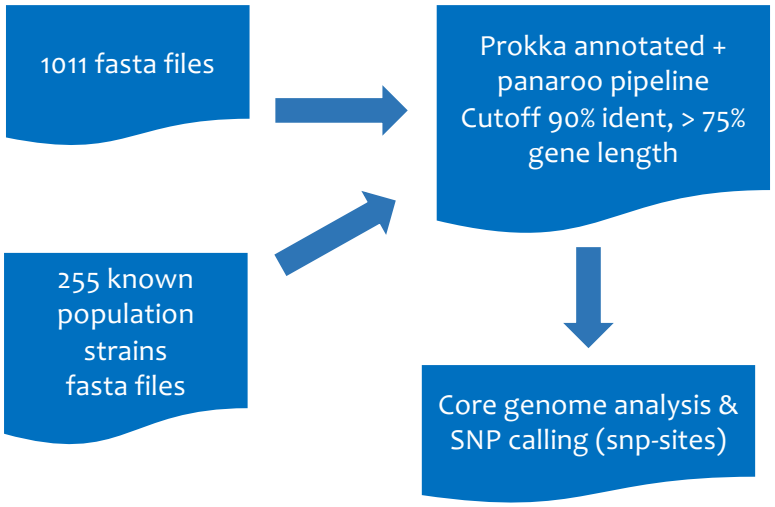
The number of core genes will approach to a constant number when N increases!

~ 800 core genes in H. pylori

https://en.wikipedia.org/wiki/Pan-genome

Focus of the *H. pylori* landmark paper

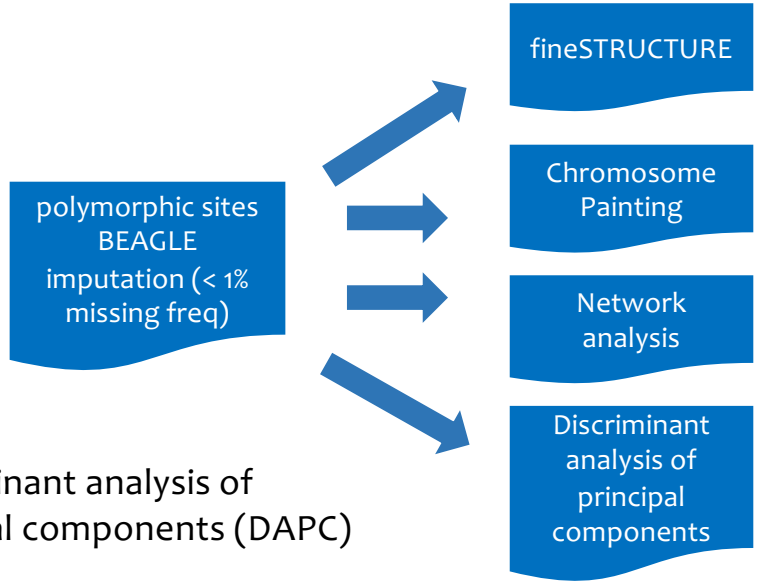
Analysis workflow



Panaroo pipeline can handle the N+1 issue. But Roary pipeline can not.

Kasia, Zilia

- Describe the dataset and genomic variation
- Map population structure
- Ancestry analysis
- Highlight interesting features of different areas



Discriminant analysis of principal components (DAPC)

DAPC

STRUCTURE vs. fineSTRUCTURE

•**Data Type:** STRUCTURE works with genotype data, while FineSTRUCTURE operates on **phased haplotype data**.

•**Scope of Analysis:** STRUCTURE is commonly used for inferring population-level admixture, while FineSTRUCTURE is more focused on fine-scale population structure and relationships.

•**Output Representation:** STRUCTURE typically presents results as bar plots, while FineSTRUCTURE outputs heatmaps and dendrograms.

In summary, STRUCTURE and FineSTRUCTURE serve similar purposes in terms of population structure analysis, but FineSTRUCTURE provides a more detailed examination of fine-scale genetic relationships by leveraging haplotype information and producing informative visualizations. The choice between the two tools often depends on the specific objectives of the study and the level of resolution required in understanding population structure.

fineSTRUCTURE Results

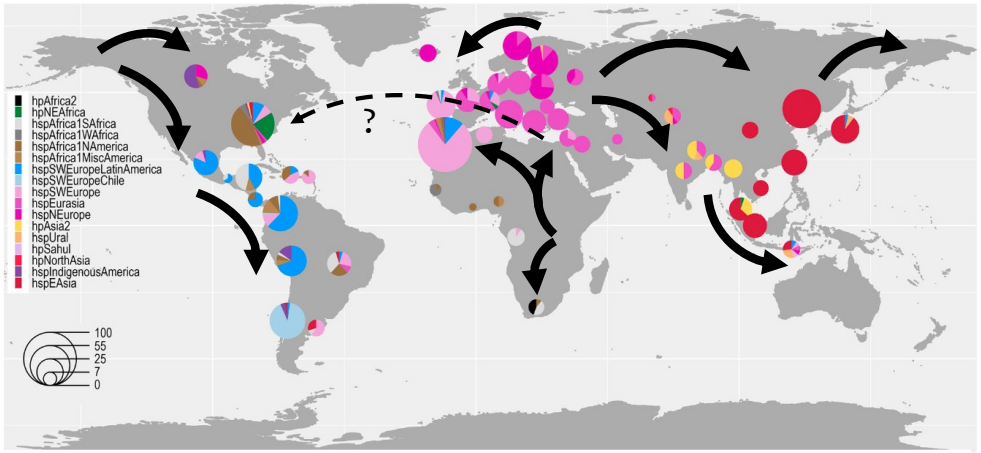
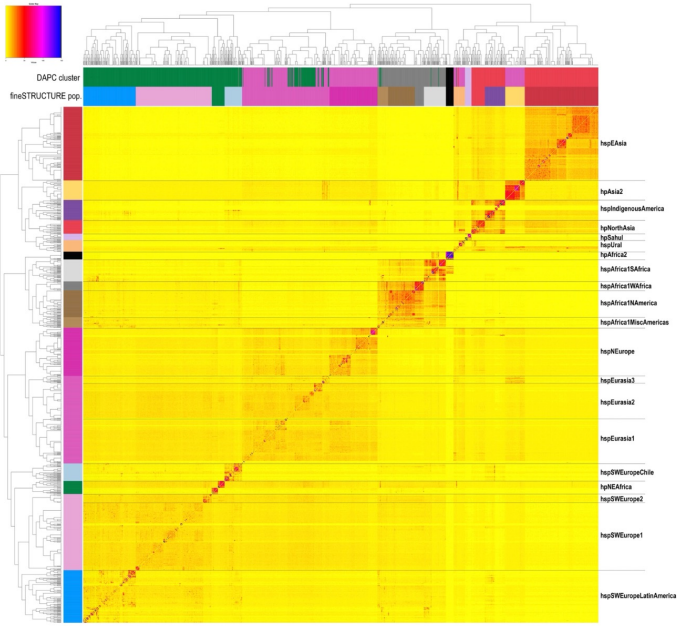


Fig. 1 | World map of *Hp*GP strain origins and population assignments. The area of each pie is proportional to the number of *Hp*GP genomes from each country and colored by the *H. pylori* population (hp) and subpopulation (hsp) as assigned by fineSTRUCTURE (Supplementary Figs. 1 and 2).



Supplementary Figure 1. Population structure of global *H. pylori* strains. The colour of each cell of the matrix indicates the expected number of DNA chunks imported from a donor genome (column) to a recipient genome (row). The inferred tree was generated by Bayesian clustering in fineSTRUCTURE. The colour bars on the top and left indicate suggested *H. pylori* population (hp) and subpopulation (hsp) as in Fig. 1, and the discriminant analysis of principal components, DAPC, K=6 cluster (Fig S3), respectively.

- Revealed four main population clusters
  (i) Southwest Europe, including Latin America and Northeast Africa
  (ii) Northern and Central Europe, Middle East, and Central Asia
  (iii) Western and Southern Africa, including Africa2 and North, South and Central America
  (iv) North, Central and East Asia, and Indigenous populations in America.

It further formed 17 main subpopulations.

- The hpEurope subpopulations span from the Atlantic coast to South Asia
- Clear differences between East and West Europe. hspNEurope further spans to South Asia.

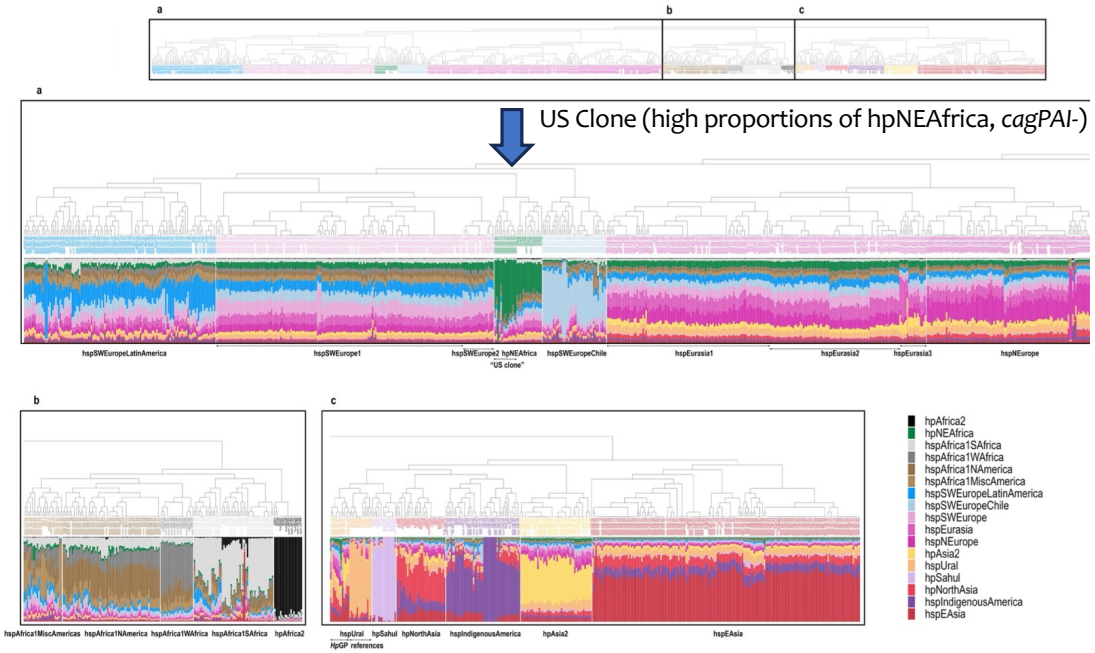
Strains from Eastern Europe and the Middle East are new in this study and have rarely been studied before.

3D plot: https://hpgp.shinyapps.io/Interactive_figures/

Kasia, Zilia

Chromosome Painting Results

Chromosome painting proportions for each genome



17 subpopulations

Kasia, Zilia

Ancestral analysis of Central Asian genomes

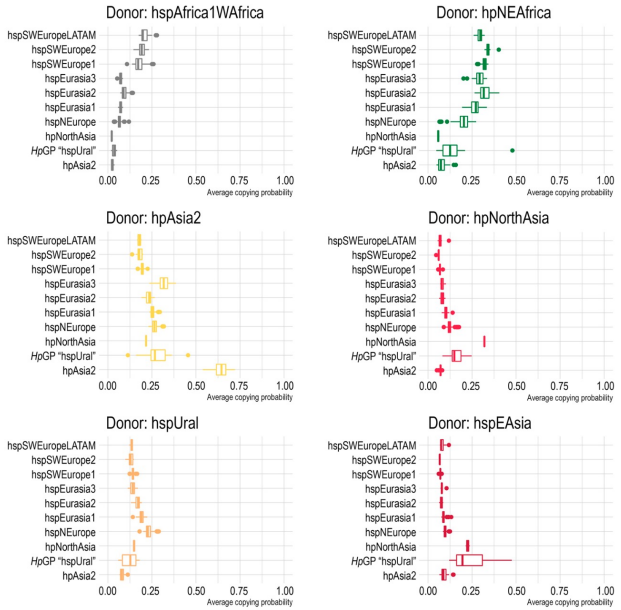
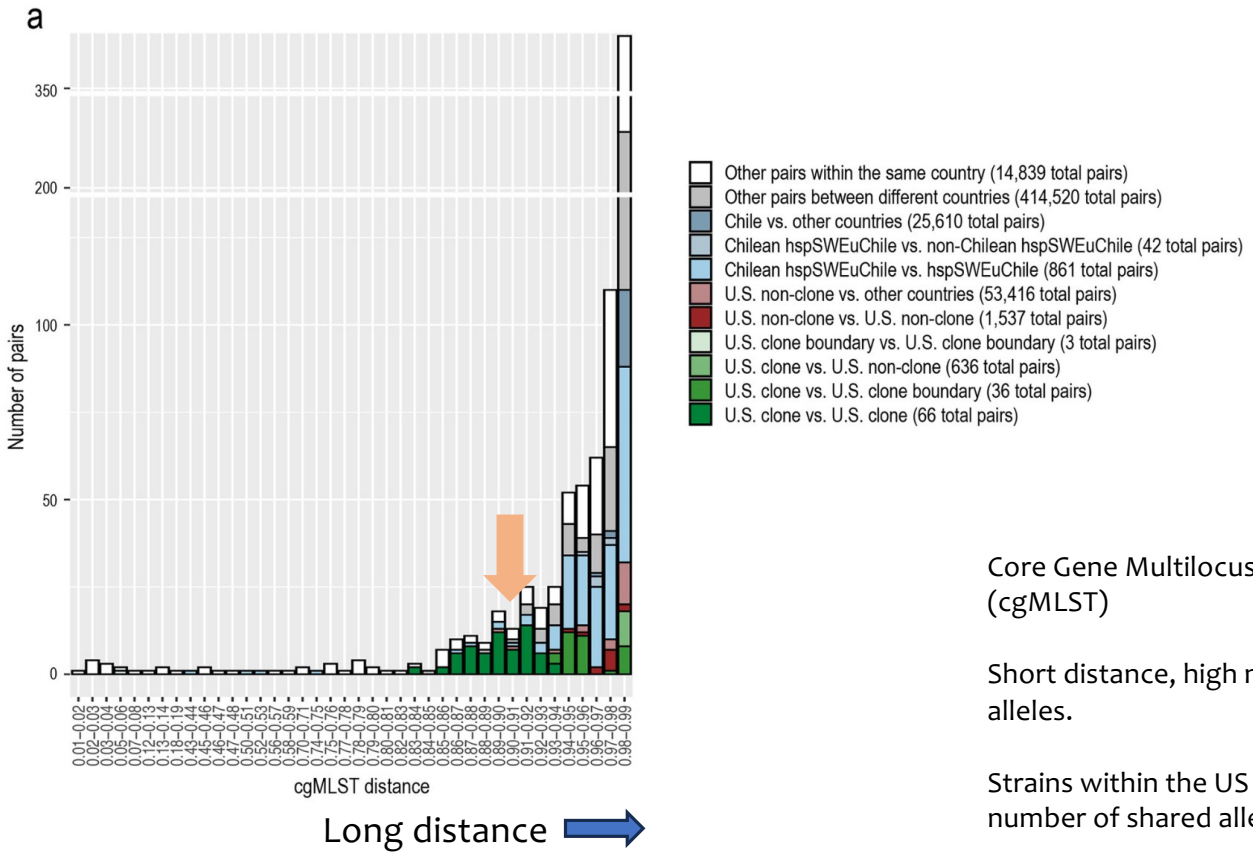


Fig. 3 | Inferred ancestral genomic contributions to the Eurasian *HpGP* genomes. Ancestral chromosome painting proportions by donor and Eurasian subpopulation. Boxplots show the median value per group, and the 25th and 75th percentiles (hinges), with whiskers extending from the hinge to the largest value no further than 1.5 × IQR (inter-quartile range) from the hinge. Data points beyond the whiskers are plotted individually. The number of genomes in each respective Eurasian population is hspSWEuropeLatinAmerica, $n = 15$; hspSWEurope2, $n = 12$; hspSWEurope1, $n = 129$; hspEurasia3, $n = 18$; hspEurasia2, $n = 76$; hspEurasia1, $n = 103$; hspNEurope, $n = 95$; hpNorthAsia, $n = 2$; *HpGP* "hspUral", $n = 10$; hpAsia2, $n = 27$.

The ancestral contributions to the **central Asian** genomes (**Eurasian**) confirmed the HpGP "hspUral" clade not to have pronounced contribution by the hspUral references but **relatively high hpAsia2, hpNorthAsia and hspEAsia painting proportions**

In-depth Analysis of the US Deep Clonal Relationships in HpGP



Core Gene Multilocus Sequence Typing (cgMLST)

Short distance, high number of shared alleles.

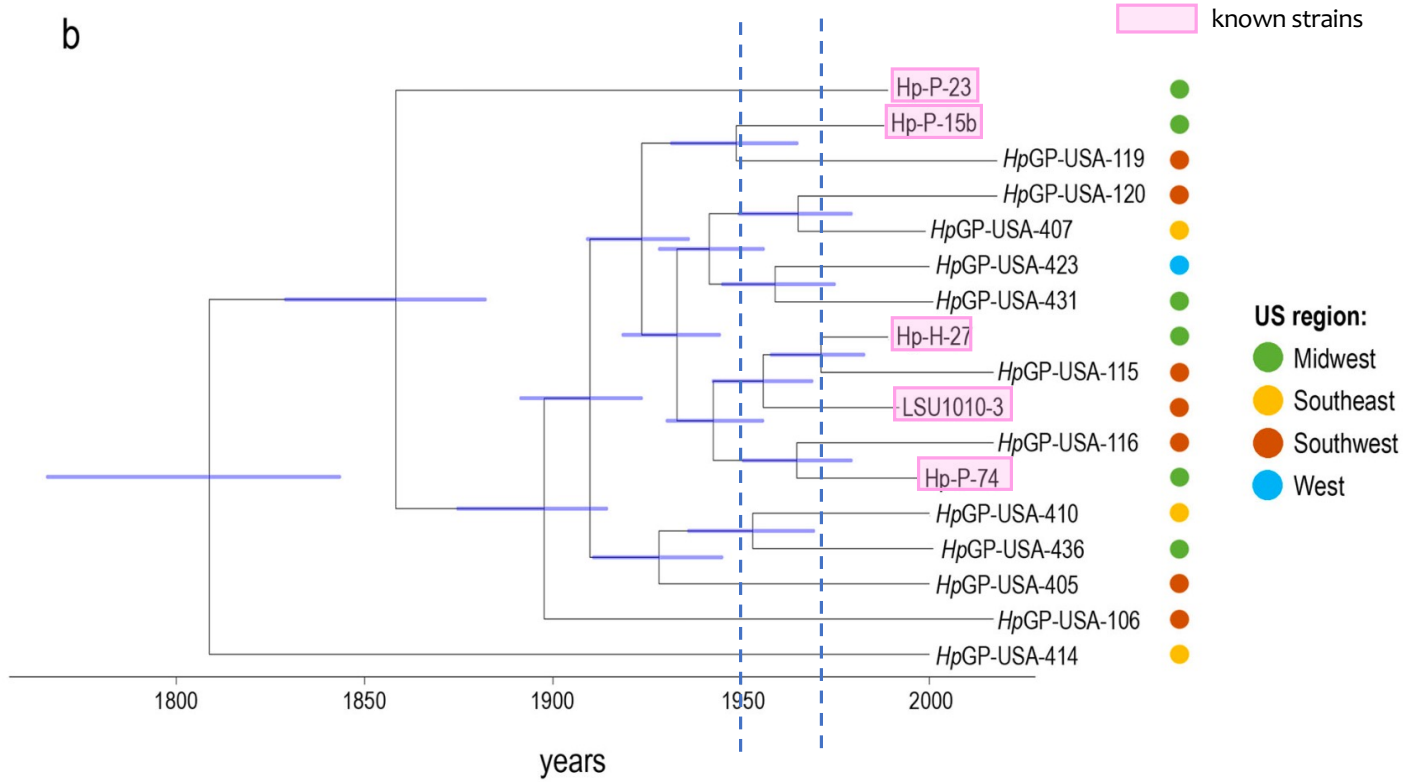
Strains within the US clone have high number of shared alleles.

Kasia, Zilia

In-depth Analysis of the US Deep Clonal Relationships in HpGP

ClonalFrameML tree

Estimated it from a common ancestral strain ~ 175 years ago in the US.



Kasia, Zilia

Summary of Population Classifications

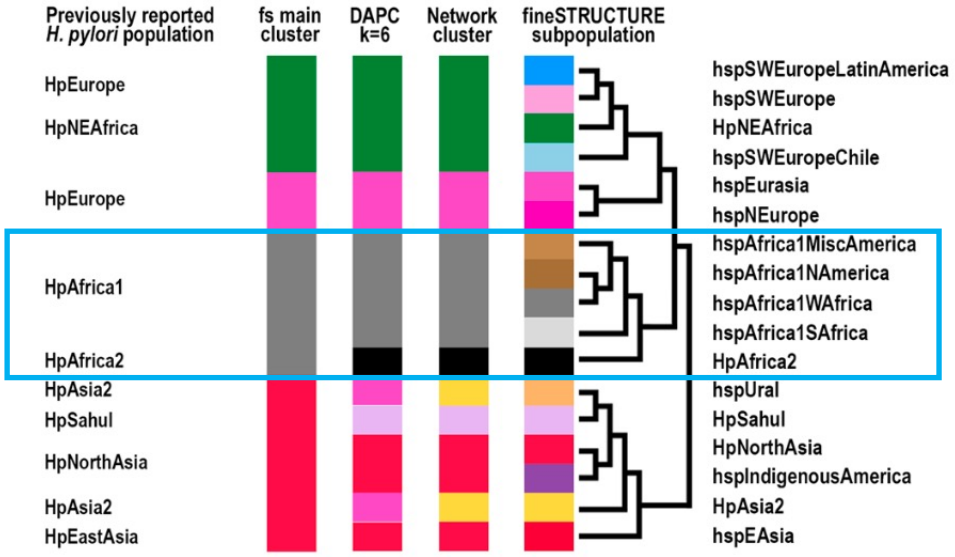
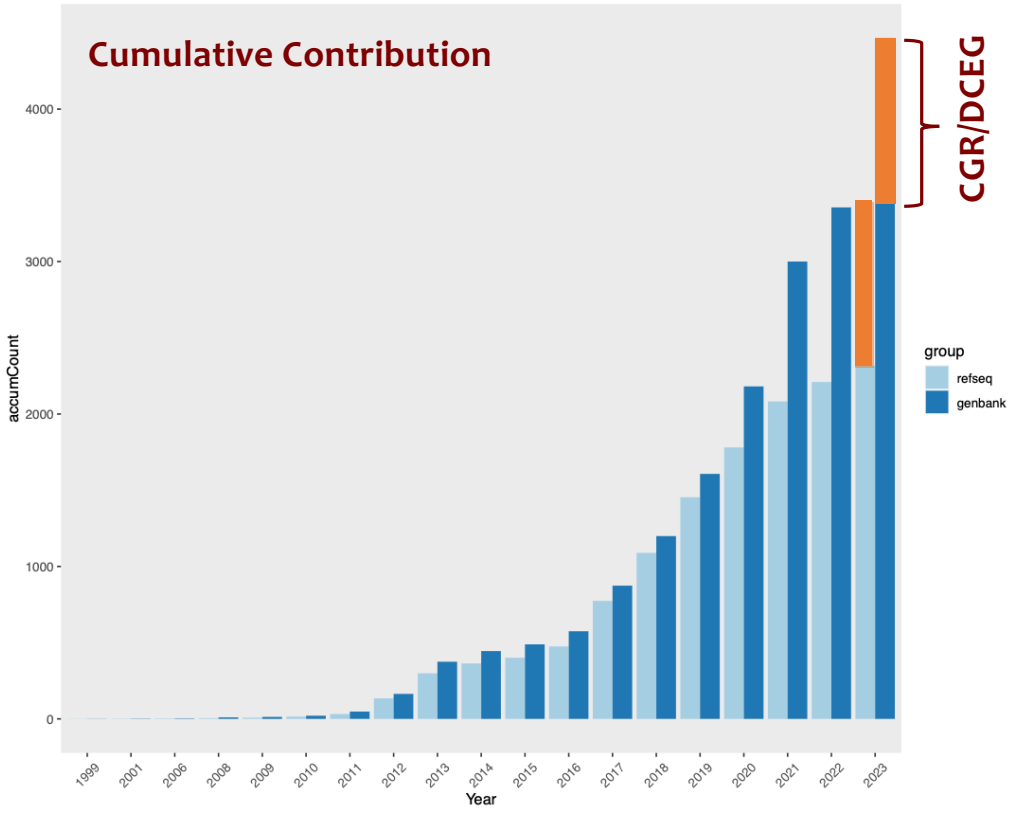
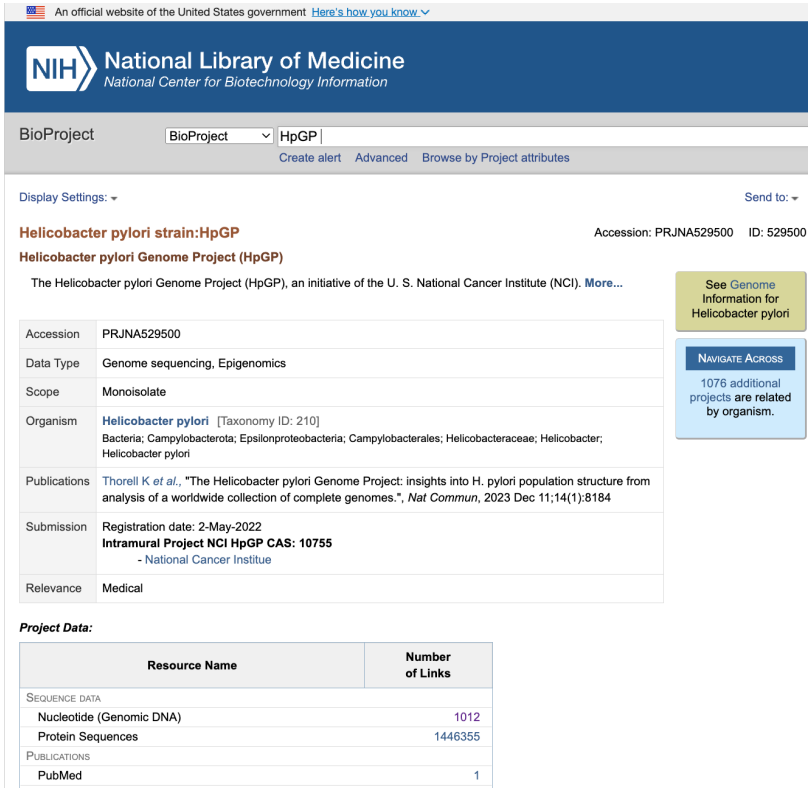


Fig. 5 | Summary of population classifications. Summary of the clustering results using the respective analyses in relation to previously reported MLST and whole genome-based *H. pylori* populations (Hp) and subpopulations (hsp). Colors are based on classifications from the fineSTRUCTURE (fs) analyses visualized in Supplementary Fig. 1, on the $K = 6$ discriminant analysis of principal components, DAPC (Supplementary Fig. 3), and the network clusters (Fig. 2). The topology of the dendrogram to the left is based on the fineSTRUCTURE hierarchical clustering of Supplementary Fig. 1.

Kasia, Zilia

Complete *H. Pylori* Genomes in GenBank (2024-01-20)





https://www.ncbi.nlm.nih.gov/bioproject/PRJNA529500

https://zenodo.org/records/10048320

Acknowledgment

2019

CGR
Sequencing team
Bioinformatics team
Wen, Kristie, Kedest, Yunhu, Belynda

DCEG
Maria Constanza Camargo
Charles Rabkin
Kai Yu and Bin Zhu

2023

FNLCR
Sequencing and Bioinformatics teams

CGR Research Analysis Support Group
Chad, Sambit, Weiyin and Xin