Strain Aware Metagenome Assembly from Short Reads with StrainXpress

Alexander Schönhuth
Bielefeld University

ICCMG, Geneva
November 16, 2023

Overview

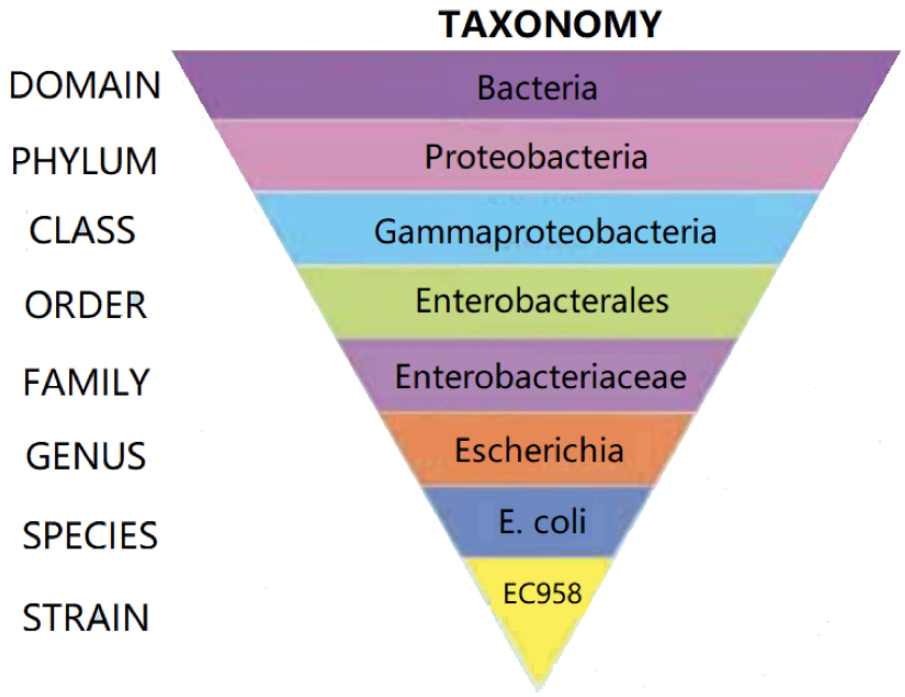
- Motivation
  - Strains matter
  - Overlap Graphs
  - Short Reads
- Methods
  - Workflow: Divide and Conquer
  - Workflow Stages
- Results
  - Simulated Data
  - Real Data
- Outlook

Motivation I: Strains Matter

Strain Level Genome Assembly

Advantages

- Clinically relevant phenotypes vary at strain level
  - E.g. some E.coli strains inflammatory, while ...
  - ... other benefit hosts by producing vitamin K
- However:
  - Species level metagenome assembly done deal
  - **Strain aware metagenome assembly most driving methodological challenge**



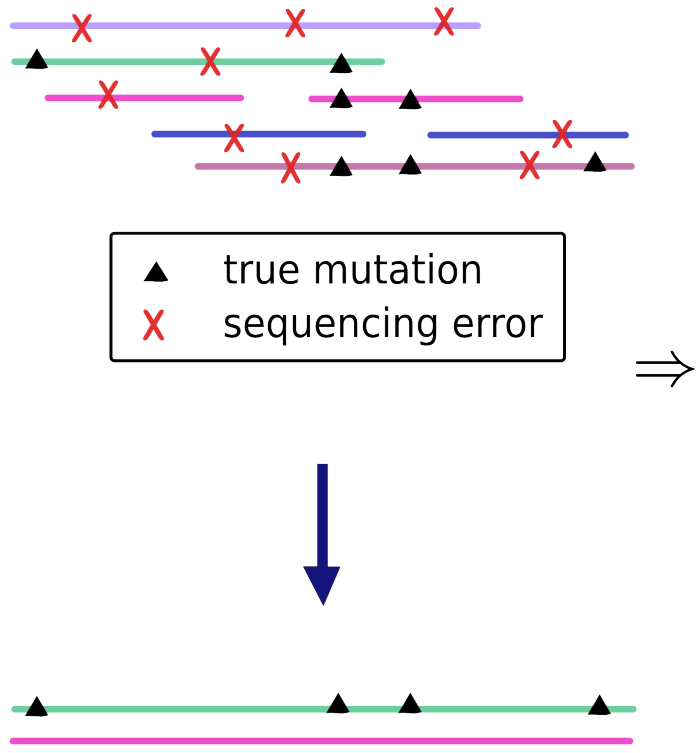
Motivation II: Overlap Graphs

Co-occuring Mutations

Challenges:

1. Connect reads/contigs from the same strain

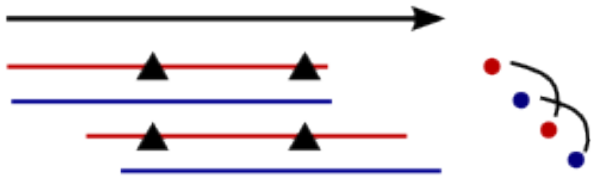
2. Distinguish sequencing errors from low-frequency mutations

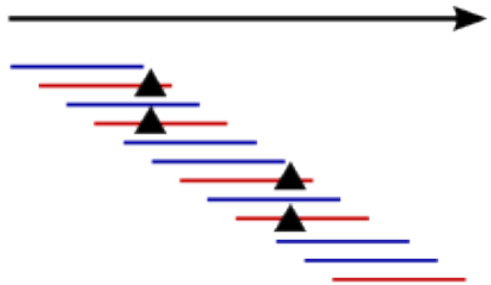


Identification of **co-occurring mutations** is key

Assembly Graphs

	Overlap graphs	De Bruijn graphs
nodes:	sequencing reads	length *k* substrings
edges:	approximate suffix-prefix overlaps	exact overlaps of length *k-1*





Overlap graphs preserve co-occurrence information

Overlap Graphs and Metagenome Assembly

Viral Quasispecies: Problem Solved

- Contig Assembly: [Baaijens et al., Gen.Res. 2017]
- Full-length: [Baaijens et al., Bioinf.2018/RECOMB 2020]

Metagenomes: Prior Work Fragmentary

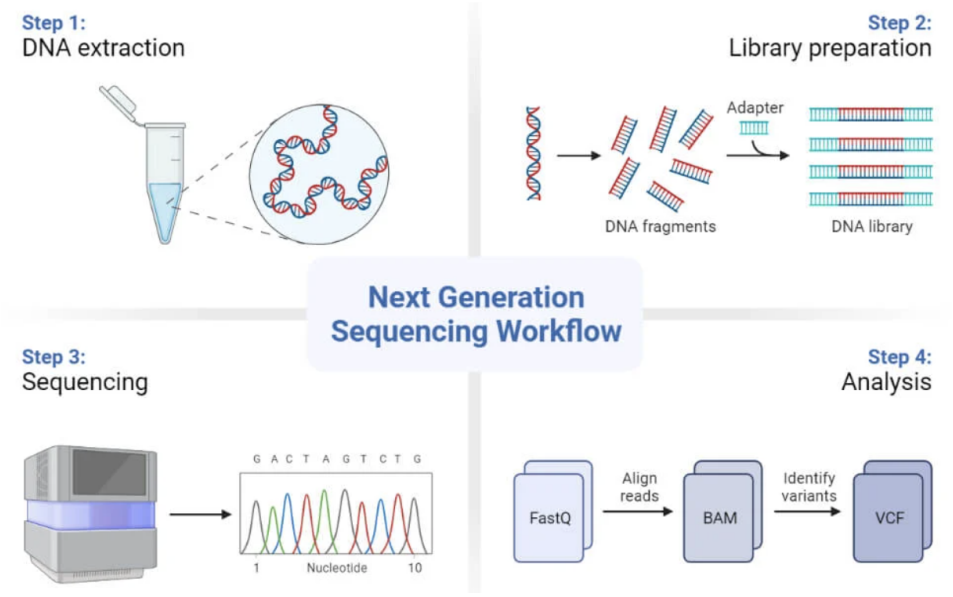
- Metagenome read clustering: [Balvert et al., Bioinformatics 2021]
- Polyploid genomes: [Baaijens et al., Bioinformatics 2019]

Summary and Approach

- Prior approaches employ DBG's, operate at species level
- *Here:* We combine existing work and add missing pieces

Motivation III: Short Reads

Next Generation Sequencing



https://microbenotes.com/next-generation-sequencing-ngs/

- Most mature, little operational burden
- Low cost, little errors, high coverage easily affordable

Use your short reads – inexpensive high-end technology!

Methods

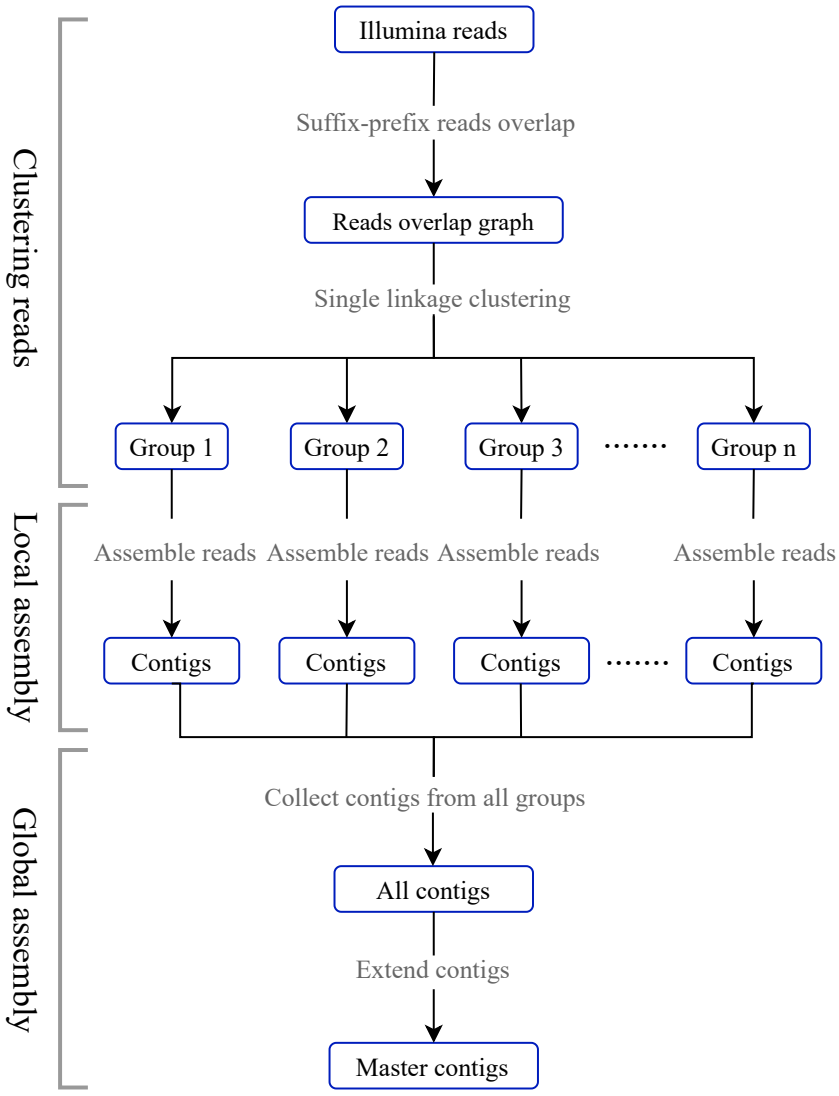
WORKFLOW

Divide

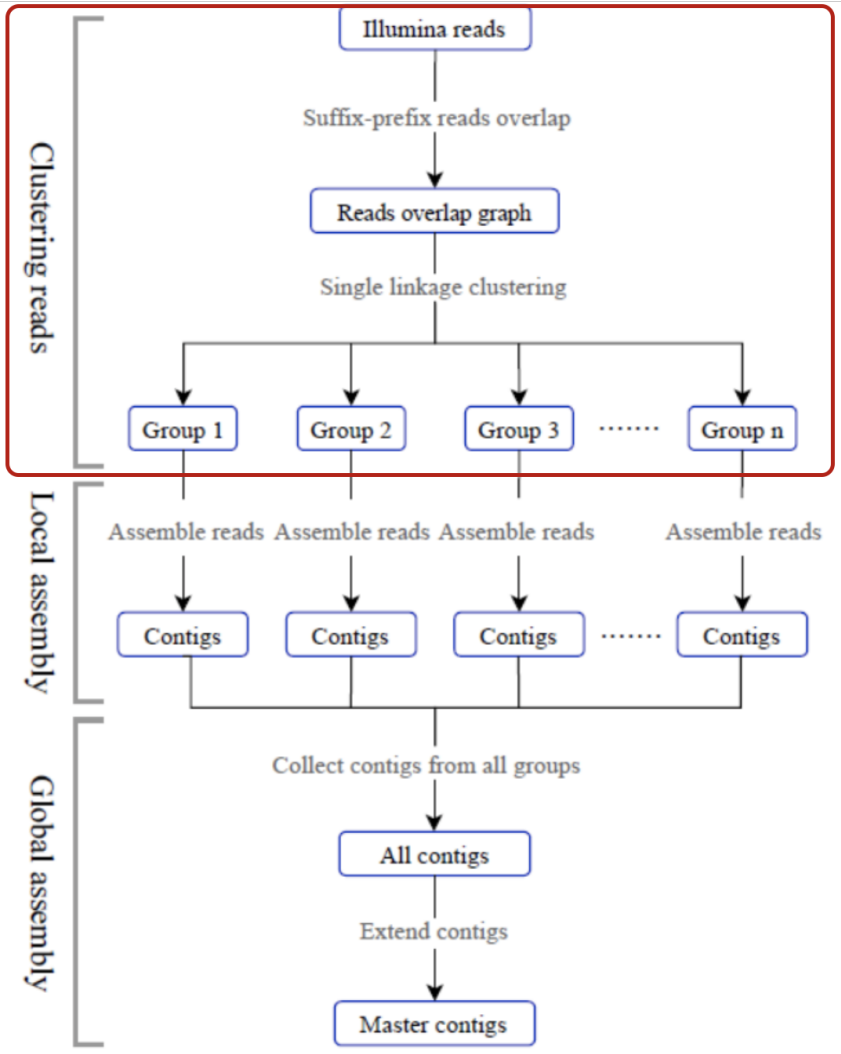
- *Clustering:* Read clusters reflect reads from same species
- *Local Assembly:* Assemble strain-specific contigs in clusters

Conquer

- *Global Assembly:* Extend contigs across clusters to optimal length

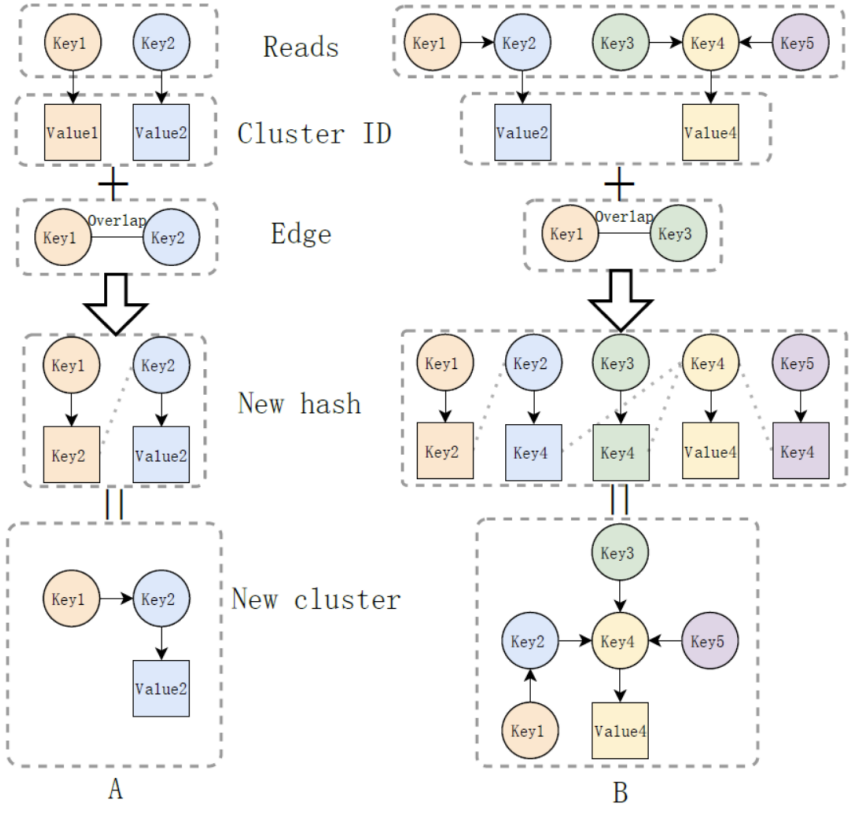


Divide: Clustering

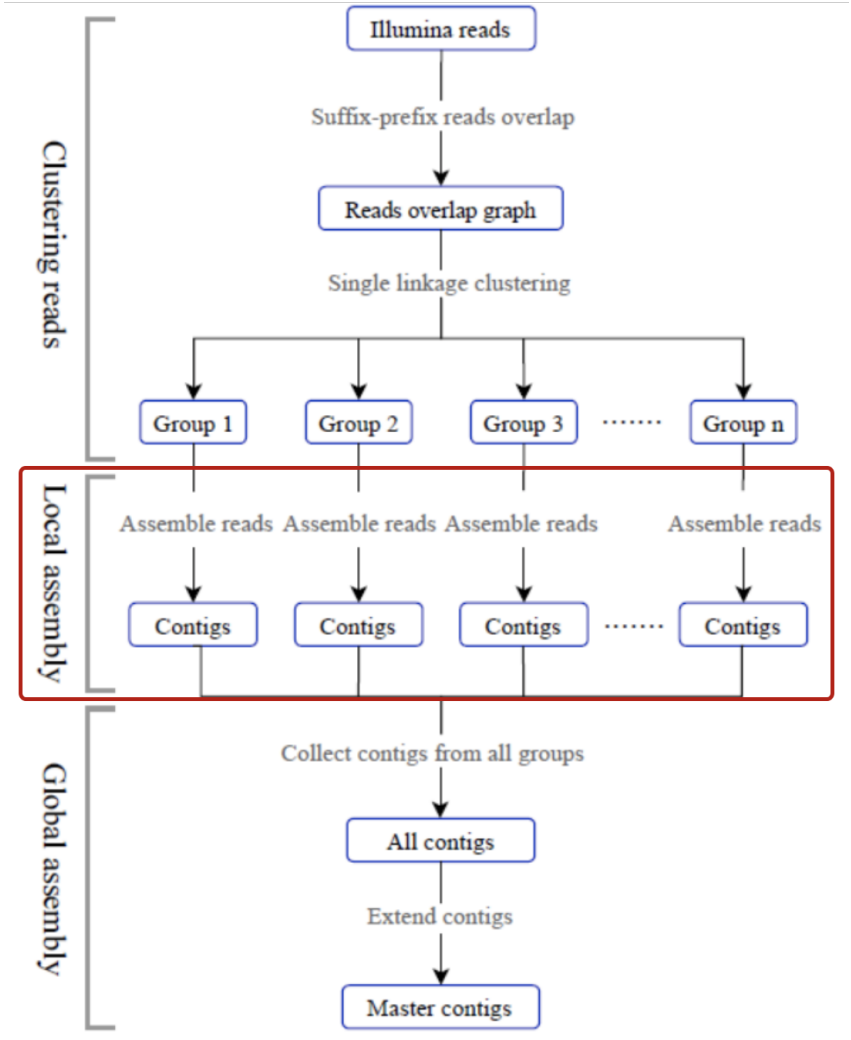


Single Linkage Clustering

- Compute overlaps between all reads
- Sort overlaps by overlap score
- Merge reads / clusters using fast hash-based algorithm
- Adopted from [Balvert et al., 2021]
- Accelerated by inexpensive algorithmic protocol, instead of machine learning based approach

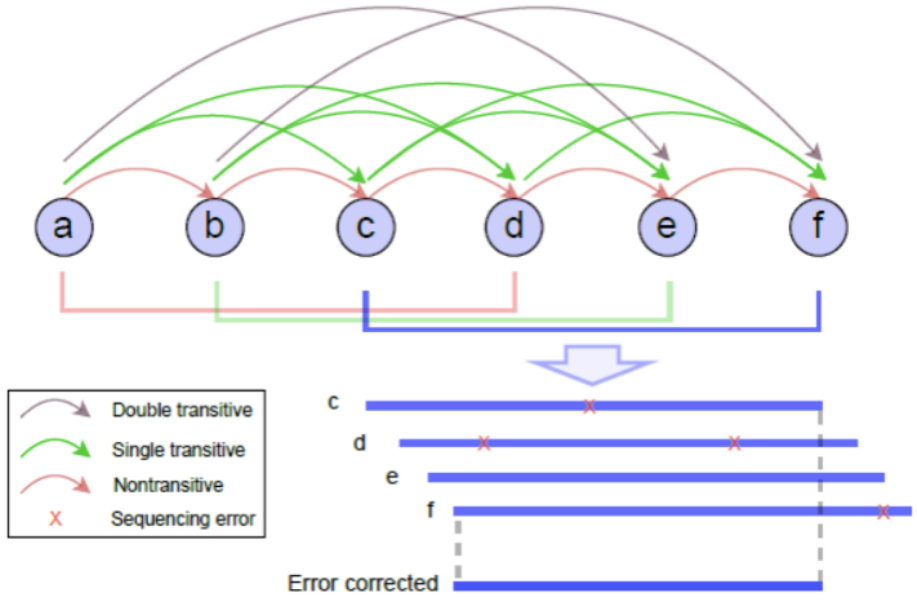


Divide: Local Assembly



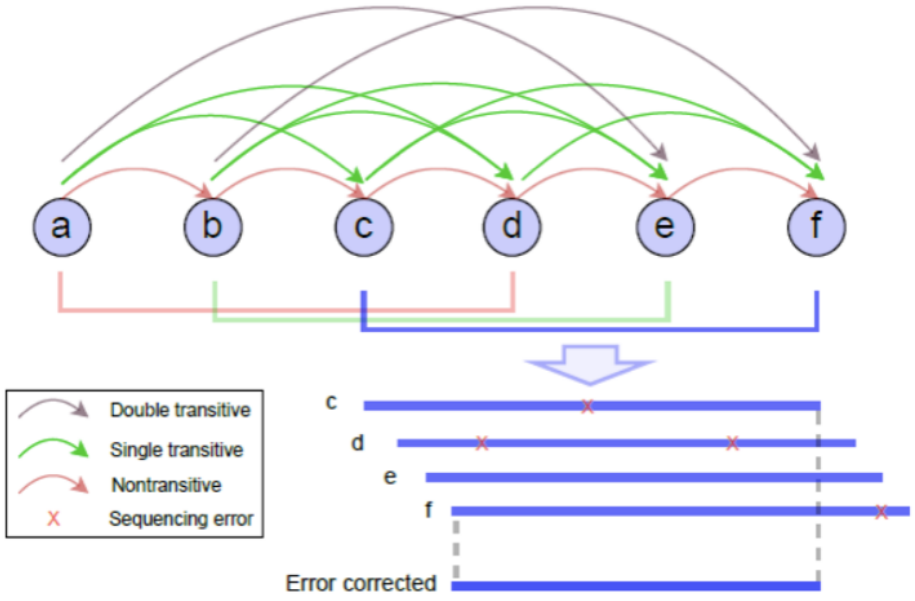
Enumerating Cliques I

- Construct overlap graph within clusters
- Adopted from [Baaijens et al., 2019]: Here *accelerated* by non-FM index based overlap strategy
- Remove transitive edges to avoid computational explosion later



Enumerating Cliques II

- Compute all maximal cliques
  ☞ fast because no transitive edges
- Remove errors, keep mutations
  ☞ in cliques mutations co-occur, but errors are isolated
- Extend reads in clique into contigs

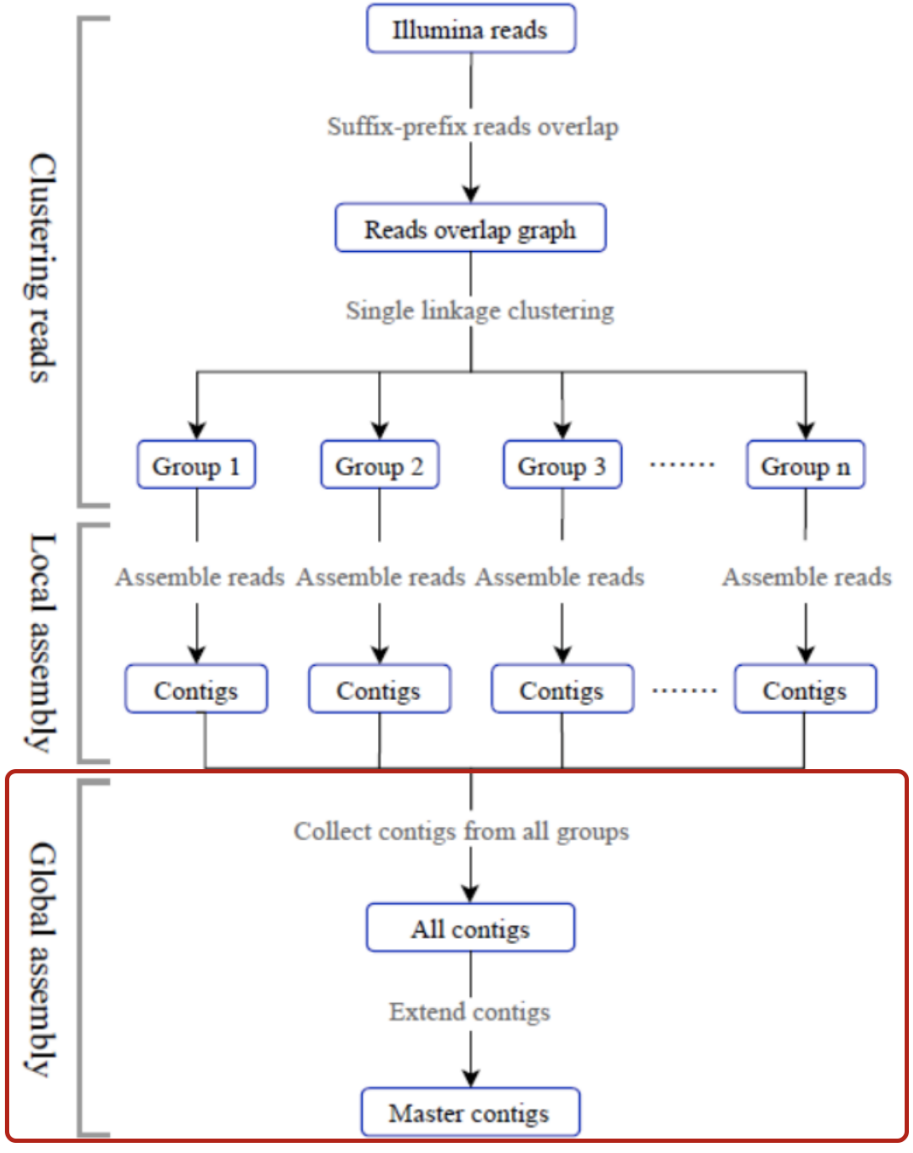


Contigs strain aware because of preservation of strain specific mutation

Conquer: Global Assembly

- *Standard Procedure:*
  - Construct overlap graph from cluster contigs
  - Remove transitive edges
  - Join "branch-less" contigs
- *Key to Success:*
  - Input contigs error-free
  - Input contigs strain-aware

Conquering is straightforward thanks to setup of Divide



Results

Data

- *Simulated Data:* Generated with CAMISIM
  - *Low Complexity:* 20 strains / 10 species / 20X per strain
  - *Medium Complexity:* 100 strains / 30 species / 20X per strain
  - *High Complexity:* 1057 strains / 376 species / 10X per strain
  - *Spike-in:* 10 Salmonella strains mixed into real gut microbiome
- *Real Data:*
  - *Bmock 12:* Mock community, 12 strains / 10 species
  - *NWC:* Natural whey starter culture, 6 strains / 3 species

Benchmark Experiments

- *Alternative Approaches* (all de Bruijn graph based)
  - IBDA-UD: [Peng et al., 2012]
  - GATB-Minia: [Chikhi & Rizk, 2012]
  - MEGAHIT: [Li et al., 2015]]
  - SPAdes: [Bankevich et al., 2012]
  - metaSPAdes: [Nurk et al., 2017]
- *Metrics* (QUAST)
  - *Genome Fraction:* Fraction of strain genomes assembled
    Quantifies strain awareness ☞ particular attention
  - *Identity:* Agreement of contigs with true sequence
  - *N50 / NGA50:* As usual, quantify contiguity
  - *Misassemblies / Errors:* As usual, quantify contig accuracy

Results: Low Complexity

Assembly	Genome fraction(%)	Identity(%)	Total assembly length	N50	NGA50	Misassembled contigs rate(%)	Error rate(%)
Low complexity data (20X)							
StrainXpress	**93.45**	**99.93**	99839190	2584	2955	0.11	0.06
IDBA-UD	62.73	99.91	54982071	9159	2381	**0.04**	0.09
GATB-Minia	73.33	99.78	67108819	6319	3240	0.25	0.09
MEGAHIT	62.69	99.71	55186304	**20395**	4356	1.05	0.15
SPAdes	74.74	99.91	58772582	5251	1749	0.09	**0.04**
metaSPAdes	66.56	99.76	62069880	18991	**6723**	0.10	0.11

Remarks:

- StrainXpress achieves approximately 20% better Genome Fraction
  ☞ StrainXpress (clearly) most strain aware approach
- Contiguity (NGA 50) similar to other approaches
  ☞ N50 of MEGAHIT/metaSPAdes offset by small Genome Fraction
  ☞ Other contigs are not longer!
- StrainXpress contigs contain little misassemblies and errors

Results: Medium Complexity

Assembly	Genome fraction(%)	Identity(%)	Total assembly length	N50	NGA50	Misassembled contigs rate(%)	Error rate(%)
Medium complexity data (20X)							
StrainXpress	**95.16**	**99.94**	465118278	1685	2173	**0.08**	**0.07**
IDBA-UD	62.01	99.81	216415403	5323	1059	0.22	0.08
XC+IDBA-UD	66.02	99.87	243646195	6107	1948	0.27	0.12
GATB-Minia	70.40	99.76	259945654	8324	3107	0.53	0.09
XC+GATB-Minia	70.55	99.77	267643882	8005	3305	0.61	0.11
MEGAHIT	62.77	99.47	225937990	3400	1011	14.16	0.38
XC+MEGAHIT	68.77	99.63	267230659	13454	**5384**	1.06	0.24
SPAdes	72.18	99.55	246604702	10966	2254	0.44	0.08
XC+SPAdes	59.05	99.63	206345396	4104	761	0.72	0.14
metaSPAdes	63.38	99.81	228257504	**19701**	4394	0.37	0.14
XC+metaSPAdes	53.99	99.71	195286660	6394	729	0.78	0.24

Remarks:

- Trends from Low Complexity Data re-established
- XC + *Method:* Run *Method* as local assembler in our workflow
  ☞ Enables faster and (sometimes) more favorable usage of *Method*

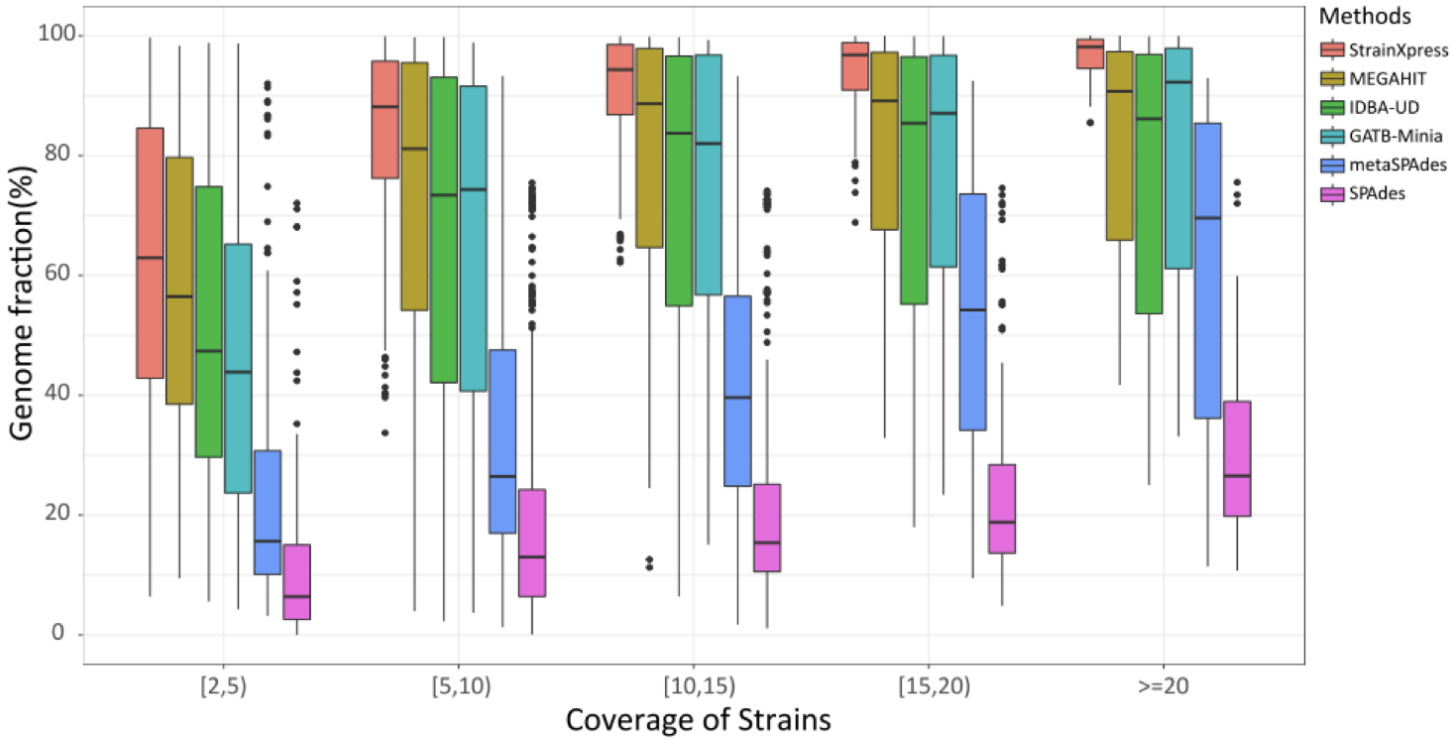
Results: High Complexity

Assembly	Genome fraction(%)	Identity(%)	Total assembly length	N50	NGA50	Misassembled contigs rate(%)	Error rate(%)
High complexity data (10X)							
StrainXpress	**84.36**	**99.88**	2278280614	1337	894	**0.14**	**0.22**
XC+IDBA-UD	70.18	99.68	2006623170	**3540**	1219	2.25	0.36
XC+GATB-Minia	68.20	99.78	1747199484	3382	794	1.07	0.25
XC+MEGAHIT	75.63	99.56	2354992154	3358	**1686**	2.14	0.87
XC+SPAdes	19.23	99.41	367171437	2542	-	3.16	0.64
XC+metaSPAdes	47.21	99.55	708596842	2613	-	1.46	0.45

Remarks:

- Explores limits of StrainXpress
  ☞ still superior, but Genome Fraction lower than 90%
- Alternative approaches crash without XC
  ☞ XC enables one to run alternative approaches on complex data
- *Overall:* All trends get re-established

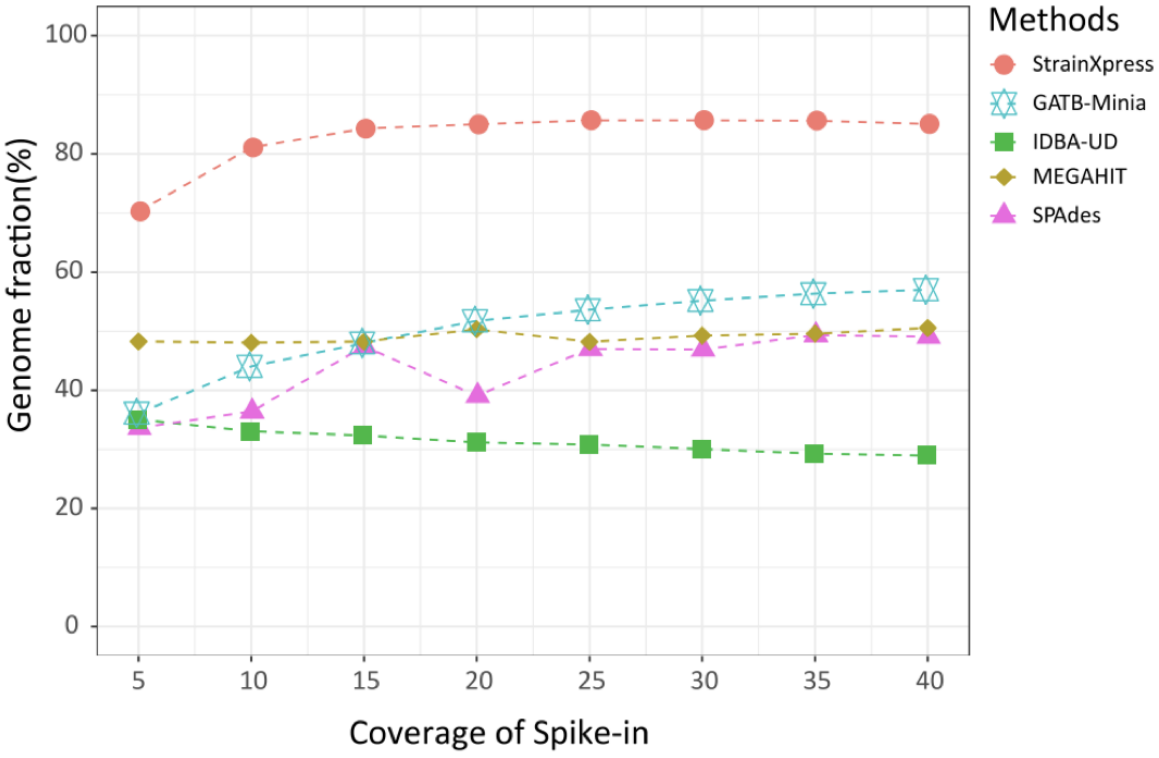
Results: Coverage



Remarks:

- Genome Fraction of StrainXpress exceeds 90% from 10X and up

Results: Spike-In Data



Remarks:

- From 15X per Salmonella strain, performance of StrainXpress stabilizes
- StrainXpress outperforms alternative approaches by at least 25%

Results: Real Data

Assembly	Genome fraction(%)	Identity(%)	Total assembly length	N50	NGA50	Misassembled contigs rate(%)	Error rate(%)
Bmock12							
StrainXpress	**99.04**	99.93	55332069	60566	65743	0.78	0.018
GATB-Minia	95.31	99.92	49058237	96537	80434	0.47	0.014
SPAdes	95.28	99.98	49055870	**189251**	**171570**	0.06	0.012
metaSPAdes	94.55	99.97	48998826	171793	155762	0.13	0.028
IDBA-UD	94.67	**99.99**	48465926	72765	60987	**0.05**	**0.006**
MEGAHIT	93.25	99.87	48637140	120129	105626	2.79	0.027
NWCs							
StrainXpress	**75.29**	99.47	8858666	1056	636	3.34	0.30
SPAdes	59.37	99.38	6083388	10160	-	2.72	0.08
metaSPAdes	57.96	**99.68**	5767394	9871	-	**1.05**	**0.05**
MEGAHIT	57.81	97.78	6141276	**14456**	-	12.52	0.16
GATB-Minia	56.78	98.78	5779411	11081	-	4.16	**0.05**
IDBA-UD	56.44	98.87	5873327	9320	-	3.97	0.07

Remarks:

- Bmock12 is easy, NWC is challenging
- By and large, trends from simulated data re-established

Results: Real Data

Strains	StrainXpress	GATB-Minia	SPAdes	IDBA-UD	MEGAHIT
Cohaesibacter sp. ES.047	98.58	97.57	98.07	97.697	97.48
Halomonas sp. HL-4	**95.54**	**69.29**	**54.39**	**56.261**	**39.07**
Halomonas sp. HL-93	**97.44**	**85.43**	**96.00**	**91.946**	**91.35**
Muricauda sp. ES.050	99.85	99.34	99.50	99.49	99.63
Micromonospora echinofusca	99.87	99.54	99.84	99.382	99.23
Marinobacter sp. LV10R510-8	**99.58**	**98.15**	**98.81**	**97.419**	**97.82**
Marinobacter sp. LV10MA510-1	**99.44**	**97.59**	**98.39**	**96.698**	**97.78**
Micromonospora echinaurantiaca	99.70	99.34	99.42	99.132	99.37
Psychrobacter sp. LV10R520-6	98.97	97.70	97.93	97.486	97.36
Propionibacteriaceae bacterium	100.00	99.98	99.99	99.948	99.96
Thioclava sp. ES.032	99.55	99.03	99.37	99.202	99.04

Remarks:

- StrainXpress only tool to distinguish between very similar strains

Conclusion / Outlook

SUMMARY

- StrainXpress only overlap graph based approach available
- StrainXpress outperforms other approaches in strain awareness
- Contigs of all tools of high quality in terms of error content

Do not forget your short reads ... they're high-end, cheap data!

Outlook: HyLight

Assembly	GF(%)	NGA50	Indels/100 kbp	Mismatches/100 kbp	N/100 kbp	MC(%)
3 *Salmonella*						
MetaPlatanus	72.25	68613	20.56	324.99	2.00	3.15
Unicycler	70.92	-	109.42	1957.53	0.00	6.88
OPERA-MS	68.43	41134	115.57	559.08	0.18	7.69
hybridSPAdes	46.22	-	35.73	816.90	0.00	1.60
HyLight	**96.03**	**351848**	0.85	**23.56**	0.00	0.19
Strainberry	-	-	-	-	-	-
StrainXpress	90.99	2645	**0.7**	59.61	0	**0.07**

Remarks:

- HyLight uses short and long reads (hybrid, lightweight assembler)
- HyLight clearly superior over all extant approaches
- In submission...

Do not forget your short reads – even when long reads are available

References

- *Paper:*

  X. Kang, X. Luo, A. Schönhuth
  StrainXpress: strain aware metagenome assembly from short reads
  Nucleic Acids Research, 50(17), e101, 2022
  `https://doi.org/10.1093/nar/gkac543`

- *Software:*

  `https://github.com/HaploKit/StrainXpress`

Thanks for your attention!