KJC 2018

An integrated simulation system for evaluating genome sequence assembly for three sequencing platforms

2018. 08. 30.

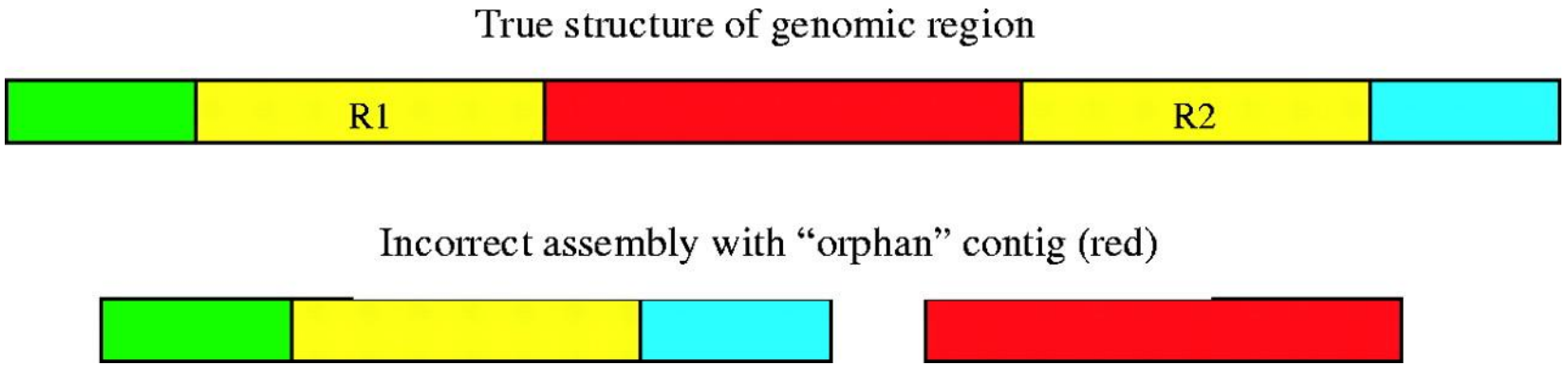
Kiejung Park,
Cheonan Campus, Sangmyung University, Cheonan, Korea

Quality of genome sequences

- 10-20 years ago – A genome project produces a (draft) genome
- Now – A genome project starts with many(draft) genomes
- Low cost of sequencing
- High quality of genome sequence?
- Errors - read, assembly, haplotying, heterozygisity, chromosomal duplication

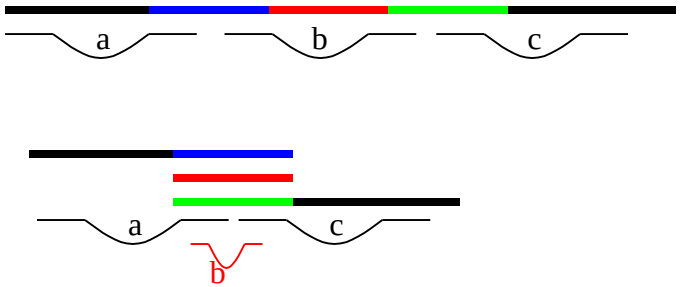
Quality of draft genomes : an example

Assemblies can collapse around repetitive sequences.

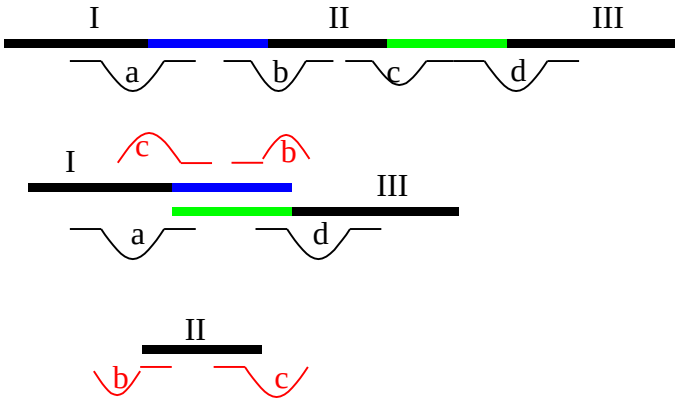


Mis-assembled repeats

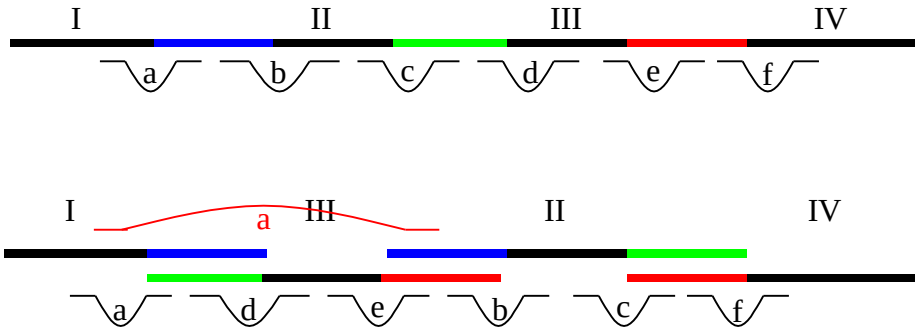
collapsed tandem



excision



rearrangement



Repeats/copies

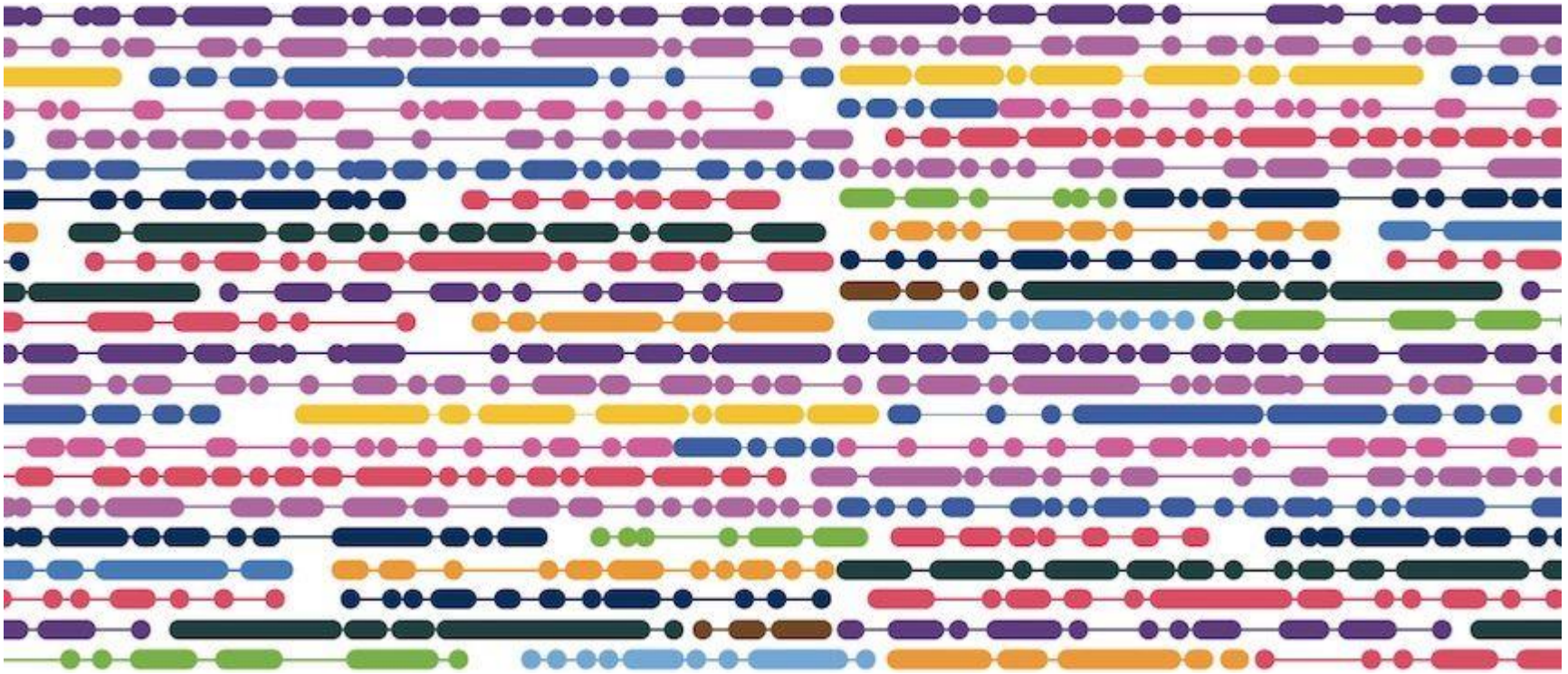
- By length
  - SNV
  - STR
  - LTR
  - CNV
- By location
  - tandom
  - Copy
  - Inversion
  - Translocation
  - Inter-/intra- chromosomal
- How to overcome?
  - (Very) long reads
  - Smart assemblers

Backgrouds and motivation

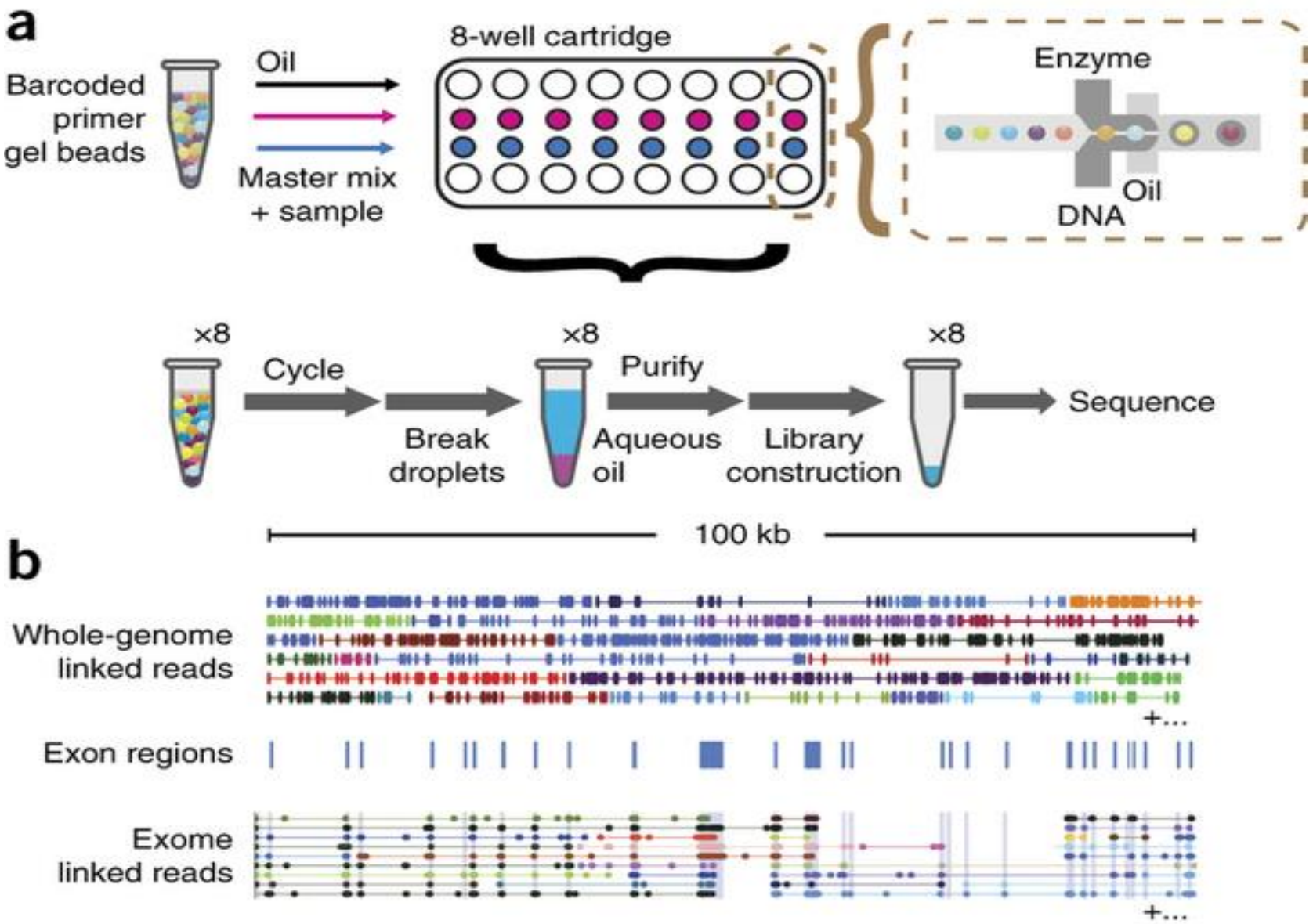
- Many types of sequencing platforms
  - Short read
  - Long read
  - Very long read
  - Linked read
- Many parameters (not standardized)
  - Cost, length, base-quality, errors
  - Data types – genome, RNA-Seq, targeted set, mapping/De novo
- Specific difficulties in marine genome analysis
  - Sample(DNA) purification
  - High heterozygosity

Linked read (seq.) technique

- A novel data type known as 'Linked-Reads' utilizes molecular barcodes to tag reads that come from the same long DNA fragment
- Library construction technique

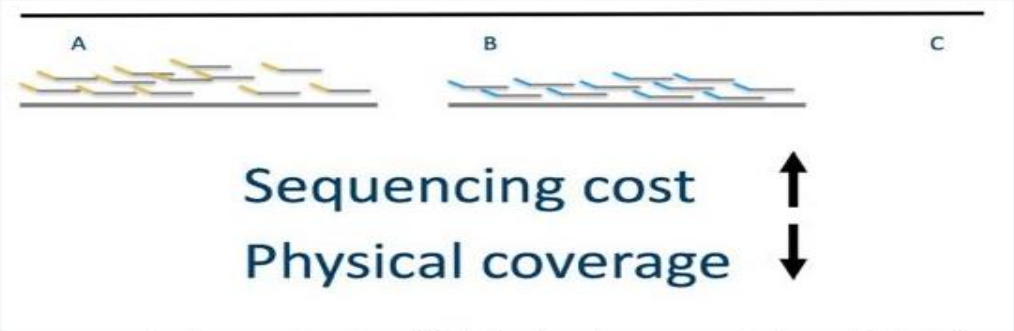


Linked reads sequencing

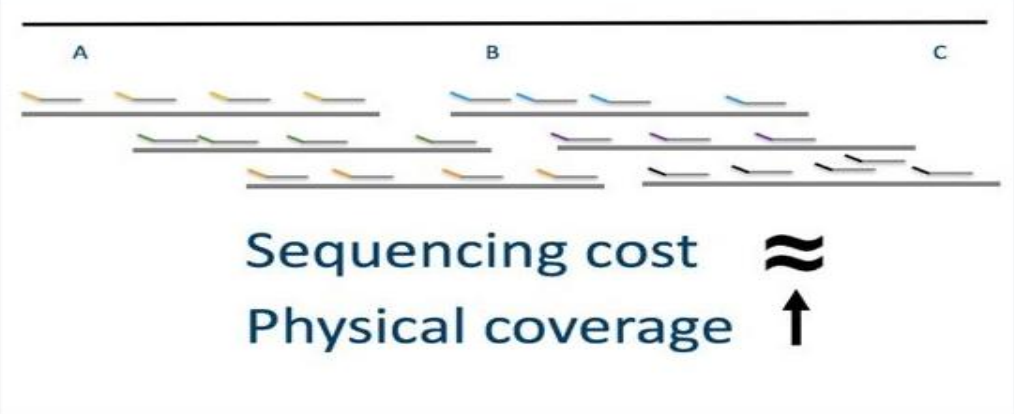


Note in the figure above, several reads (grey lines) are generated from this long input molecule and they each contain the same barcode (gold line). This allows you to deduce that these reads came from the same molecule.

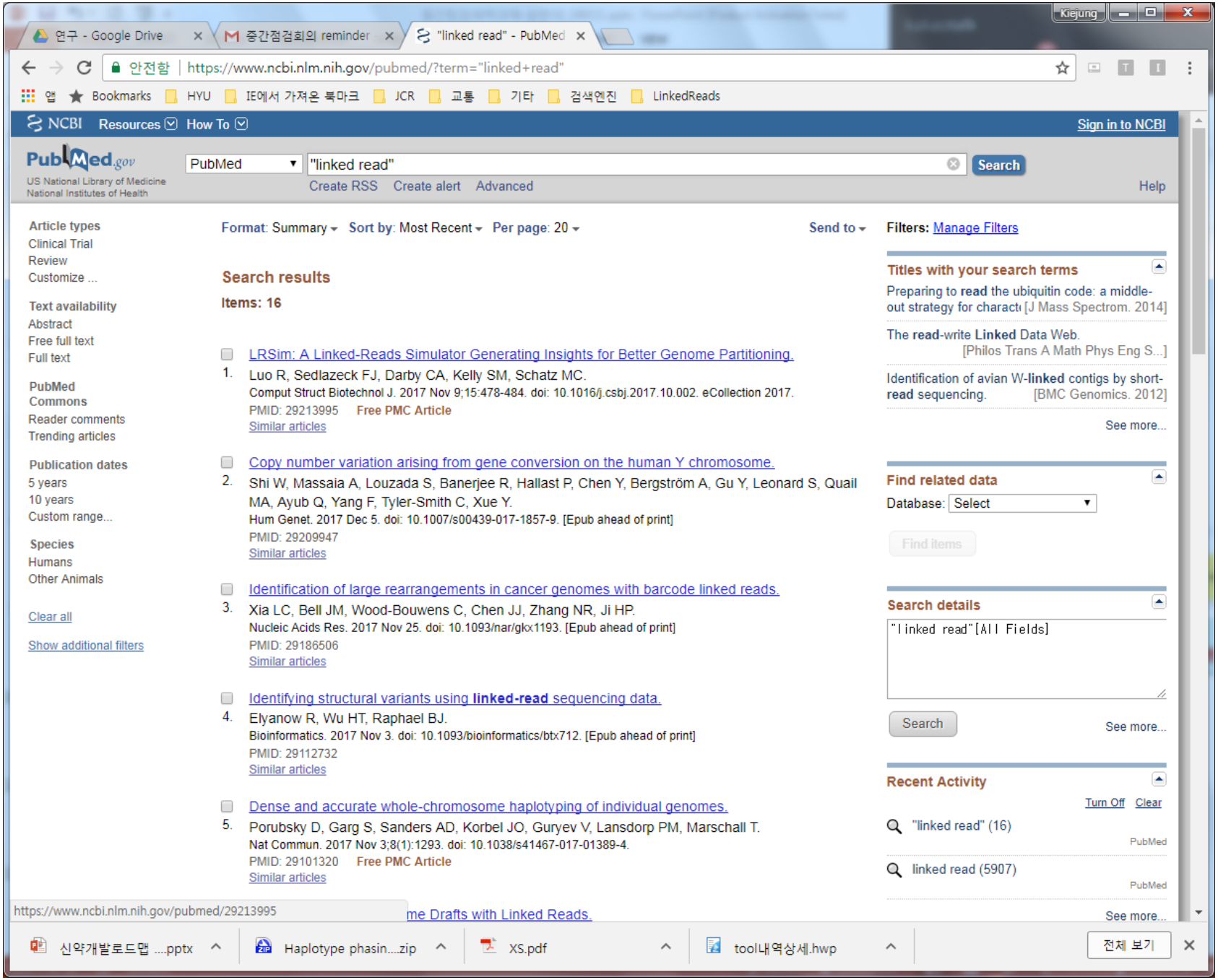
Many people wonder why we don't fully saturate the molecule with barcoded reads, to make a synthetic long read. The synthetic long read approach increases sequencing cost and typically means that you have overall less physical coverage for equivalent sequence coverage and cost.



In the image above, we see two long molecules with lots of read coverage, but we still lack the ability to link the three loci (A, B and C).



By sequencing less deeply, we can incorporate more long molecules into the system for equivalent sequence


연구 - Google Drive
중간점검회의 reminder
"linked read" - PubMed
Kiejung
안전함 | https://www.ncbi.nlm.nih.gov/pubmed/?term="linked+read"
앱 Bookmarks HYU IE에서 가져온 북마크 JCR 교통 기타 검색엔진 LinkedReads
NCBI Resources How To
Sign in to NCBI
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
"linked read"
Search
Create RSS Create alert Advanced
Help
Article types
Clinical Trial
Review
Customize ...
Text availability
Abstract
Free full text
Full text
PubMed Commons
Reader comments
Trending articles
Publication dates
5 years
10 years
Custom range...
Species
Humans
Other Animals
Clear all
Show additional filters
Format: Summary
Sort by: Most Recent
Per page: 20
Send to
Search results
Items: 16
LRSim: A Linked-Reads Simulator Generating Insights for Better Genome Partitioning.
1. Luo R, Sedlazeck FJ, Darby CA, Kelly SM, Schatz MC.
Comput Struct Biotechnol J. 2017 Nov 9;15:478-484. doi: 10.1016/j.csbj.2017.10.002. eCollection 2017.
PMID: 29213995 Free PMC Article
Similar articles
Copy number variation arising from gene conversion on the human Y chromosome.
2. Shi W, Massaia A, Louzada S, Banerjee R, Hallast P, Chen Y, Bergström A, Gu Y, Leonard S, Quail MA, Ayub Q, Yang F, Tyler-Smith C, Xue Y.
Hum Genet. 2017 Dec 5. doi: 10.1007/s00439-017-1857-9. [Epub ahead of print]
PMID: 29209947
Similar articles
Identification of large rearrangements in cancer genomes with barcode linked reads.
3. Xia LC, Bell JM, Wood-Bouwens C, Chen JJ, Zhang NR, Ji HP.
Nucleic Acids Res. 2017 Nov 25. doi: 10.1093/nar/gkx1193. [Epub ahead of print]
PMID: 29186506
Similar articles
Identifying structural variants using linked-read sequencing data.
4. Elyanow R, Wu HT, Raphael BJ.
Bioinformatics. 2017 Nov 3. doi: 10.1093/bioinformatics/btx712. [Epub ahead of print]
PMID: 29112732
Similar articles
Dense and accurate whole-chromosome haplotyping of individual genomes.
5. Porubsky D, Garg S, Sanders AD, Korbel JO, Guryev V, Lansdorp PM, Marschall T.
Nat Commun. 2017 Nov 3;8(1):1293. doi: 10.1038/s41467-017-01389-4.
PMID: 29101320 Free PMC Article
Similar articles
https://www.ncbi.nlm.nih.gov/pubmed/29213995
me Drafts with Linked Reads.
Filters: Manage Filters
Titles with your search terms
Preparing to read the ubiquitin code: a middle-out strategy for charact [J Mass Spectrom. 2014]
The read-write Linked Data Web. [Philos Trans A Math Phys Eng S...]
Identification of avian W-linked contigs by short-read sequencing. [BMC Genomics. 2012]
See more...
Find related data
Database: Select
Find items
Search details
"linked read"[All Fields]
Search
See more...
Recent Activity
Turn Off Clear
"linked read" (16)
PubMed
linked read (5907)
PubMed
See more...
신약개발로드맵 ....pptx
Haplotype phasin....zip
XS.pdf
tool내역상세.hwp
전체 보기

Greer *et al. Genome Medicine* (2017) 9:57
DOI 10.1186/s13073-017-0447-8

Genome Medicine

RESEARCH Open Access

Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases

Stephanie U. Greer[1], Lincoln D. Nadauld[1,2], Billy T. Lau[3], Jiamin Chen[1], Christina Wood-Bouwens[1], James M. Ford[1], Calvin J. Kuo[1,4] and Hanlee P. Ji[1,3*]

Abstract

Background: Genome rearrangements are critical oncogenic driver events in many malignancies. However, the identification and resolution of the structure of cancer genomic rearrangements remain challenging even with whole genome sequencing.

Methods: To identify oncogenic genomic rearrangements and resolve their structure, we analyzed linked read sequencing. This approach relies on a microfluidic droplet technology to produce libraries derived from single, high molecular weight DNA molecules, 50 kb in size or greater. After sequencing, the barcoded sequence reads provide long range genomic information, identify individual high molecular weight DNA molecules, determine the haplotype context of genetic variants that occur across contiguous megabase-length segments of the genome and delineate the structure of complex rearrangements. We applied linked read sequencing of whole genomes to the analysis of a set of synchronous metastatic diffuse gastric cancers that occurred in the same individual.

Results: When comparing metastatic sites, our analysis implicated a complex somatic rearrangement that was present in the metastatic tumor. The oncogenic event associated with the identified complex rearrangement resulted in an amplification of the known cancer driver gene *FGFR2*. With further investigation using these linked read data, the *FGFR2* copy number alteration was determined to be a deletion-inversion motif that underwent tandem duplication, with unique breakpoints in each metastasis. Using a three-dimensional organoid tissue model, we functionally validated the metastatic potential of an *FGFR2* amplification in gastric cancer.

11. Eslami Rasekh M, Chiatante G, Miroballo M, Tang J, Ventura M, Amemiya CT, Eichler EE, Antonacci F, Alkan C.
BMC Genomics. 2017 Jan 10;18(1):65. doi: 10.1186/s12864-016-3444-1.
PMID: 28073353 **Free PMC Article**
Similar articles

HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies.
12. Edge P, Bafna V, Bansal V.
Genome Res. 2017 May;27(5):801-812. doi: 10.1101/gr.213462.116. Epub 2016 Dec 9.
PMID: 27940952 **Free PMC Article**
Similar articles

Universal Haplotype-Based Noninvasive Prenatal Testing for Single Gene Diseases.
13. Hui WW, Jiang P, Tong YK, Lee WS, Cheng YK, New MI, Kadir RA, Chan KC, Leung TY, Lo YM, Chiu RW.
Clin Chem. 2017 Feb;63(2):513-524. doi: 10.1373/clinchem.2016.268375. Epub 2016 Dec 8.
PMID: 27932412
Similar articles

Assembly of the Complete Sitka Spruce Chloroplast Genome Using 10X Genomics' GemCode Sequencing Data.
14. Coombe L, Warren RL, Jackman SD, Yang C, Vandervalk BP, Moore RA, Pleasance S, Coope RJ, Bohlmann J, Holt RA, Jones SJ, Birol I.
PLoS One. 2016 Sep 15;11(9):e0163059. doi: 10.1371/journal.pone.0163059. eCollection 2016.
PMID: 27632164 **Free PMC Article**
Similar articles

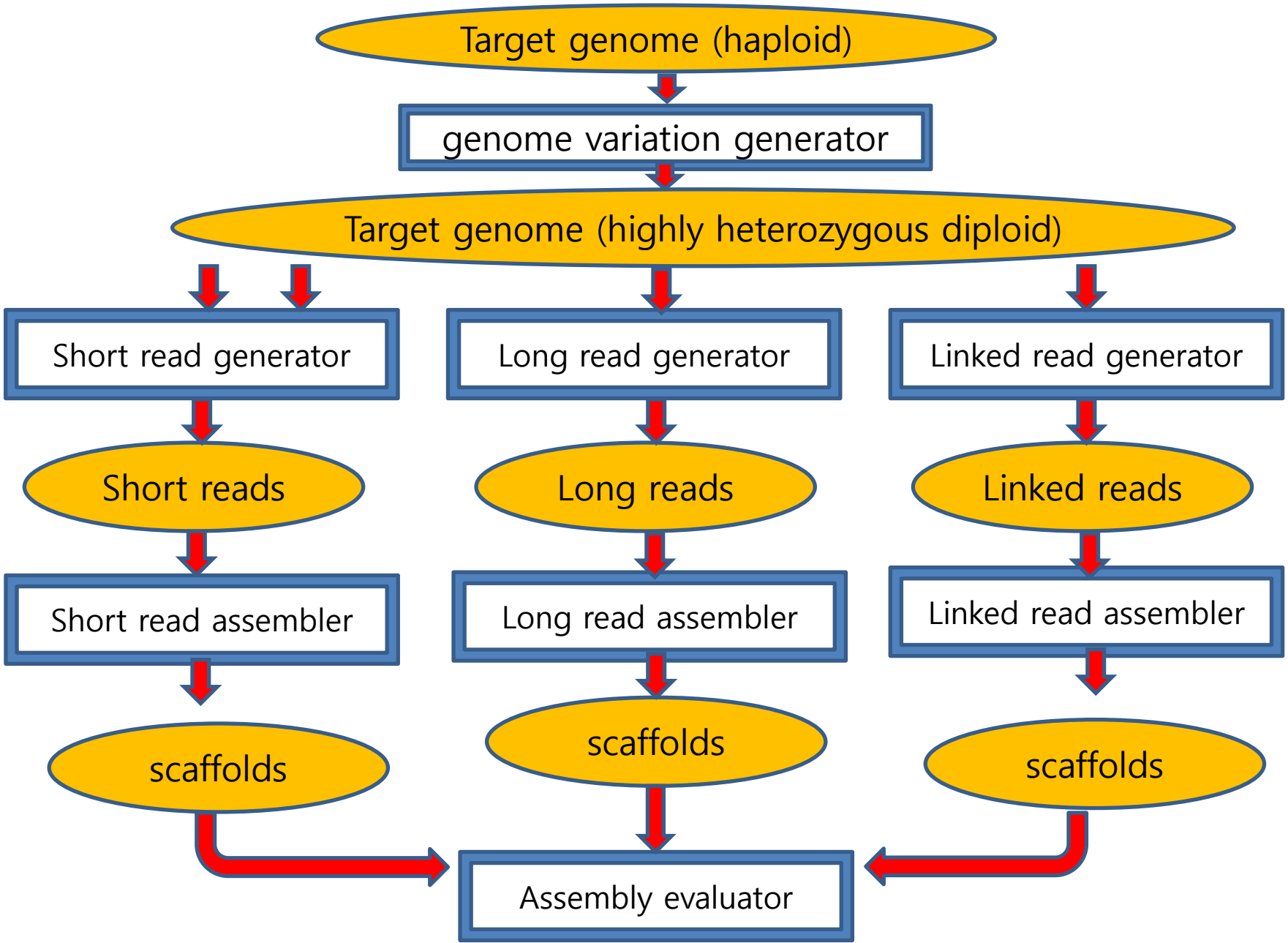
A hybrid approach for de novo human genome sequence assembly and phasing.
15. Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastie AR, Marks P, Lee J, Chu C, Lin C, Džakula Ž, Cao H, Schlebusch SA, Giorda K, Schnall-Levin M, Wall JD, Kwok PY.
Nat Methods. 2016 Jul;13(7):587-90. doi: 10.1038/nmeth.3865. Epub 2016 May 9.
PMID: 27159086 **Free PMC Article**
Similar articles

Haplotyping germline and cancer genomes with high-throughput **linked-read** sequencing.
16. Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, Mudivarti PA, Wyatt PW, Bharadwaj R, Makarewicz AJ, Li Y, Belgrader P, Price AD, Lowe AJ, Marks P, Vurens GM, Hardenbol P, Montesclaros L, Luo M, Greenfield L, Wong A, Birch DE, Short SW, Bjornson KP, Patel P, Hopmans ES, Wood C, Kaur S, Lockwood GK, Stafford D, Delaney JP, Wu I, Ordonez HS, Grimes SM, Greer S, Lee JY, Belhocine K, Giorda KM, Heaton WH, McDermott GP, Bent ZW, Meschi F, Kondov NO, Wilson R, Bernate JA, Gauby S, Kindwall A, Bermejo C, Fehr AN, Chan A, Saxonov S, Ness KD, Hindson BJ, Ji HP.
Nat Biotechnol. 2016 Mar;34(3):303-11. doi: 10.1038/nbt.3432. Epub 2016 Feb 1.

Considerations for linked read techniques

- Lower cost than long read platforms
- Better assembly than short read platforms -> effective in phasing haplotype like long read platforms
- Problems – not verified yet, very difficult to implement good assemblers -> Currently seems practical by benchmarking
- Needs for evaluation for applying to marine genomes
- Good for marines genomes with high heterozygosity

Pipeline for evaluating seq. platforms



Chromosome variation generator

- With single chromosome
- base substitution
- deletion
- transposition
- inversion
- duplication and tandom duplication

Read generators

- Template DNA → (error options) → FASTQ (w/QV)
- XS – short reads generator
- SimLoRD – long reads generator
- LRSim - linked reads generator

- Standardization – error options

Assemblers

- Diploid genomes
- SOAPdenovo – short reads assembler
- Falcon assembler – long reads assembler
- Supernova - linked reads assembler

- Standardization – error/coverage options

Assembly evaluator

- Block homology against original genome sequences – BLAST, BLAT, .. (error range)
- Block match count
- Block linkage count

- What is better assembly?

9 (mutational) events for variation

- % option
- Random position, random order
- Intra chromosomal

substitution 5
short_insertion 0.1
short_deletion 0.1
insertion 0.01
deletion 0.01
transposition 0.0001
inversion 0.0001
duplication 0.0001
tandom_duplication 0.0001

mkgenvar - Chromosome variation generator

kjpark@whale-portal:~/seqsim

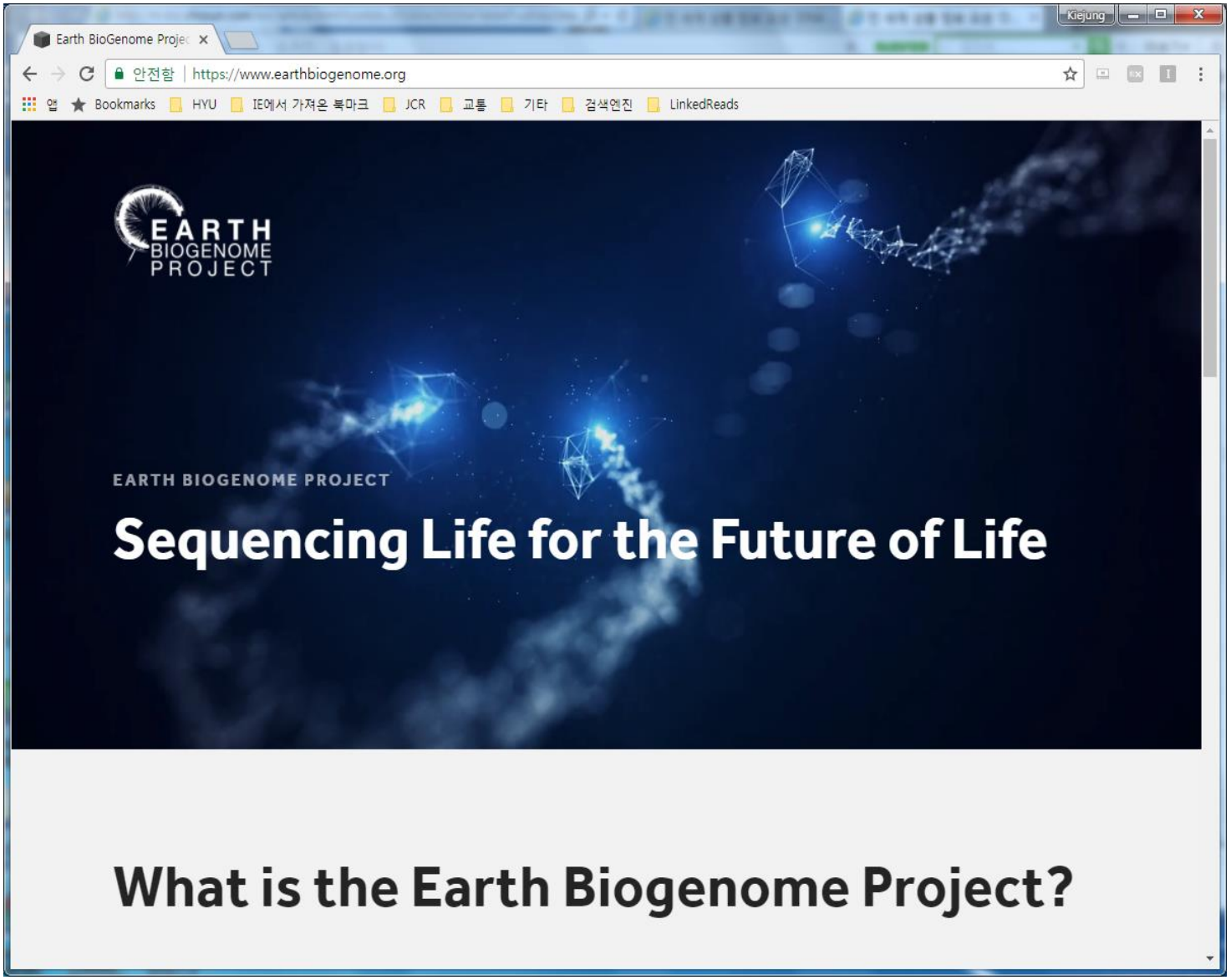
```
[kjpark@whale-portal seqsim]$ mkgenvar
Usage : mkgenvar <input_file> <var_option_file> <output_file> <var_event_file>
Can not open <input_file>
[kjpark@whale-portal seqsim]$ mkgenvar 1M.fna s5_id01_ID001_others00001.opt 1M_var.fna
1M.var | more
Check : pass arg
Seq : >NC_019949.1 Mycoplasma cynos C142 complete genome(50)

seqlen = 998117
input seq len = 998117
total var eventnum = 50712
event 0/50712
subst (551379): T -> A
event 1/50712
shortdelete (801511): A ->
shortdelete: event 2/50712
subst (399474): T -> A
event 3/50712
subst (666894): A -> G
event 4/50712
subst (307206): T -> A
event 5/50712
subst (546165): A -> C
event 6/50712
subst (156007): C -> A
event 7/50712
subst (318223): A -> C
event 8/50712
subst (69542): G -> T
event 9/50712
subst (350045): T -> C
event 10/50712
subst (853364): G -> C
event 11/50712
subst (241988): T -> C
event 12/50712
subst (258447): T -> G
event 13/50712
subst (954959): G -> T
event 14/50712
subst (995796): A -> C
event 15/50712
shortdelete (265065): T ->
shortdelete: event 16/50712
```

Futher works

- Precise/practical evaluator
- Standardization of analysis parameters among platforms
- More practical genome generator – multi-chromosomal, inter-chromosomal, CNV,…
- Web interface

What is EBP?


Earth BioGenome Proje
안전함 | https://www.earthbiogenome.org
Bookmarks HYU IE에서 가져온 북마크 JCR 교통 기타 검색엔진 LinkedReads
EARTH BIOGENOME PROJECT
EARTH BIOGENOME PROJECT
Sequencing Life for the Future of Life
What is the Earth Biogenome Project?

EBP – comprehensive sequencing of earth genomes

Earth BioGenome Proje

https://phys.org/news/2018-04-earth-biogenome-aims-sequence-dna.html

Bookmarks HYU IE에서 가져온 북마크 JCR 교통 기타 검색엔진 LinkedReads

This website uses cookies to ensure you get the best experience on our website. More info

PHYS.ORG Nanotechnology Physics Earth Astronomy & Space Technology Chemistry Biology Other Sciences

search

Home » Biology » Evolution » April 23, 2018

Earth BioGenome Project aims to sequence DNA from all complex life on Earth

April 23, 2018, UC Davis

The Earth BioGenome Project aims to sequence all eukaryotic species. This superkingdom of life includes all organisms except bacteria and archaea. Credit: Mirhee Lee

An international consortium of scientists is proposing what is arguably the most ambitious project in the history of biology: sequencing the DNA of all known eukaryotic species on Earth.