Parallel distributed genome analysis pipeline and large-scale genome data researches in KISTI

Junehawk Lee, PhD
juneh@kisti.re.kr

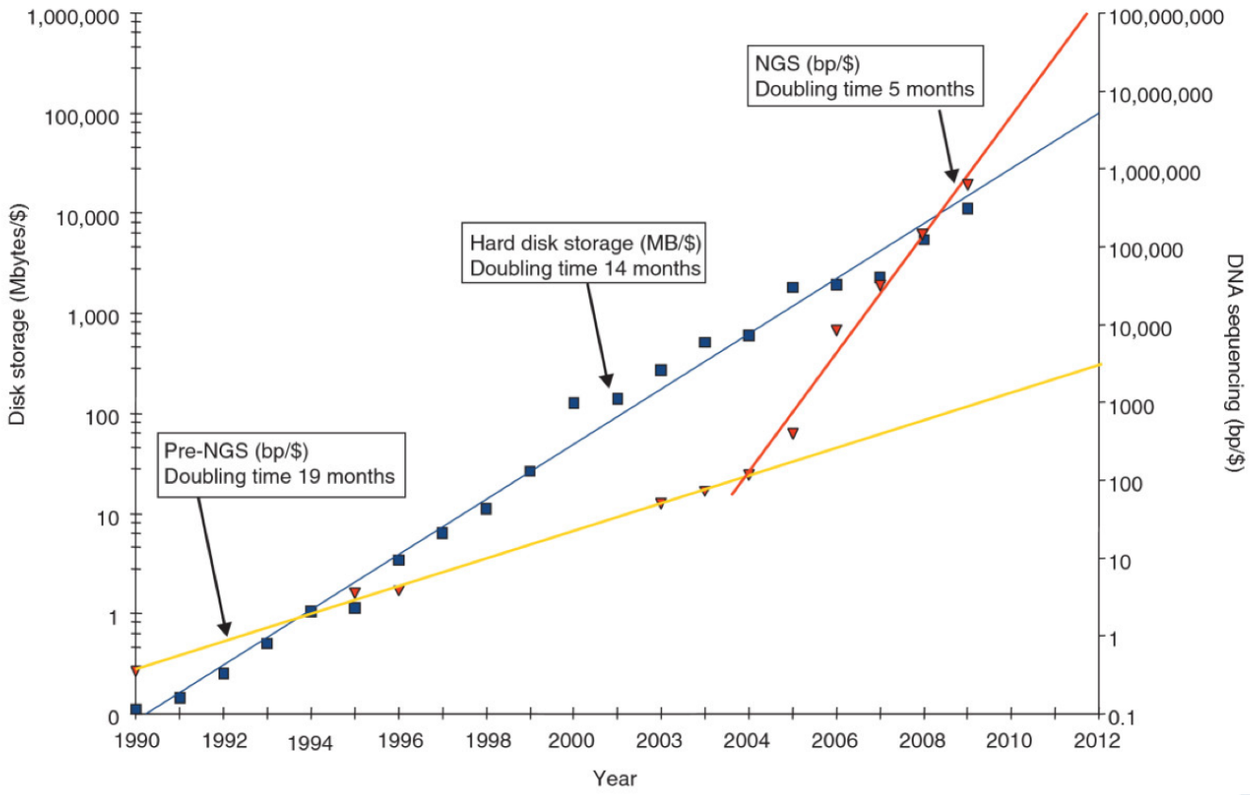
Center for Supercomputing Applications,
Korea Institute of Science and Technology Information

Contents

- Genome sequence analysis
- Parallel distributed genome analysis pipeline
- Genomics projects in KISTI
- Conclusion

Speed of genome sequence data generation

- While the time required to produce sequence data is rapidly decreasing, the time for analyzing sequence data is not improving
    - NovaSeq 6000 - 6 Tb data/20 billion reads in 2 days
- Sequence analysis can be the next bottleneck
    - Disk storage, computing costs are not decreasing fast enough



Analysis speed matters…

Science Translational Medicine Home News Journals Topics Careers

Advertisement

Log in | My account | Contac

SHARE

RESEARCH ARTICLE | DIAGNOSTICS

Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive Care Units

Carol Jean Saunders[1,2,3,4,5,*], Neil Andrew Miller[1,2,4,*], Sarah Elizabeth Soden[1,2,4,*], Darrell Lee Dinwiddie[1,2,3,4,5,*], Aaron N…

+ See all authors and affiliations

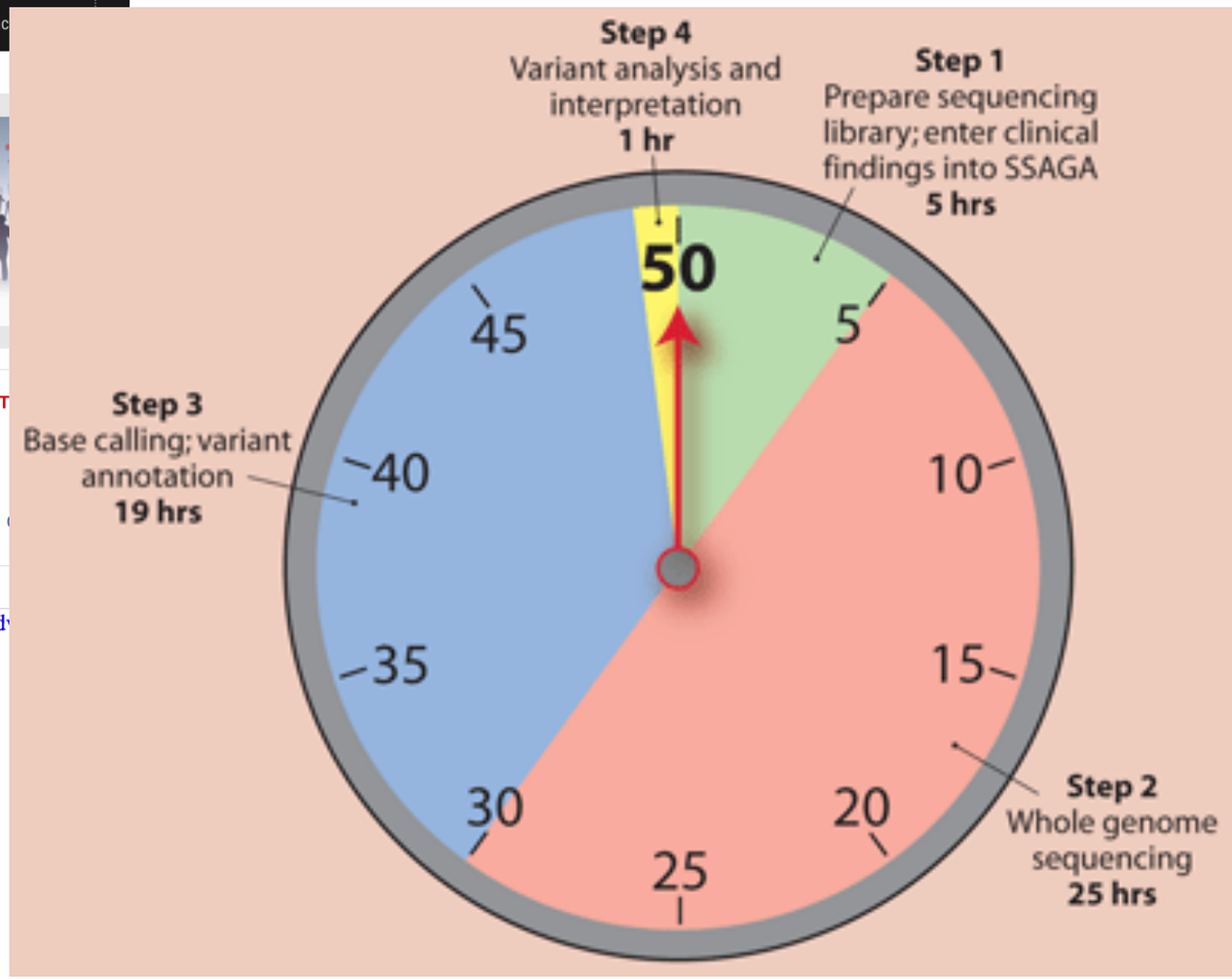
Science Translational Medicine 03 Oct 2012:
Vol. 4, Issue 154, pp. 154ra135
DOI: 10.1126/scitranslmed.3004041

Article | Figures & Data | Info & Metrics | eLetters | PDF

You are currently viewing the abstract. View Full Text

Abstract

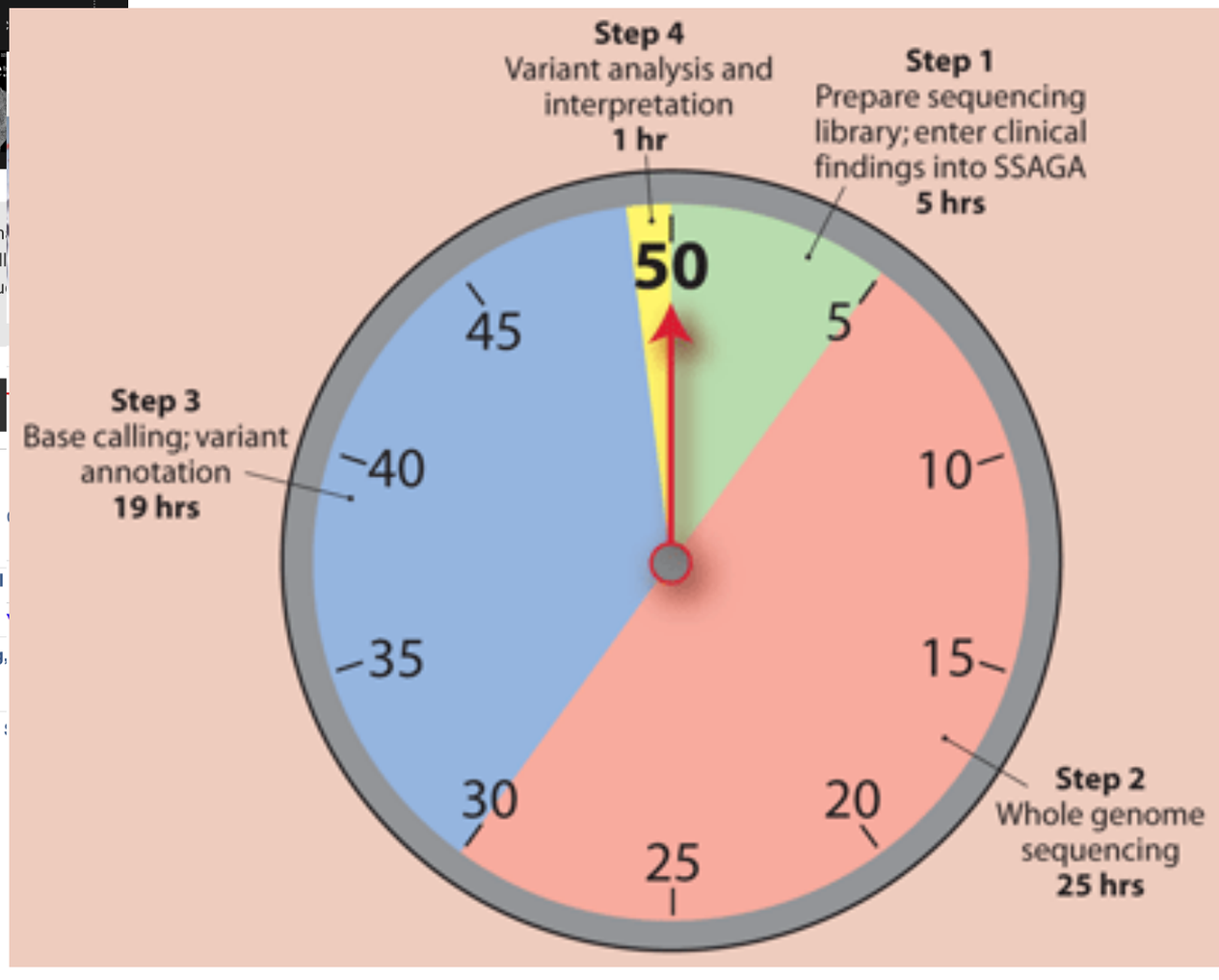
Monogenic diseases are frequent causes of neonatal morbidity and mortality, and disease presentations are often undifferentiated at birth. More than 3500 monogenic diseases have



Saunders et. al., Sci Transl Med 2012

- Emergency genomics
- Tailormade drug using mutation pattern of patients
- The time for the analysis can be further improved

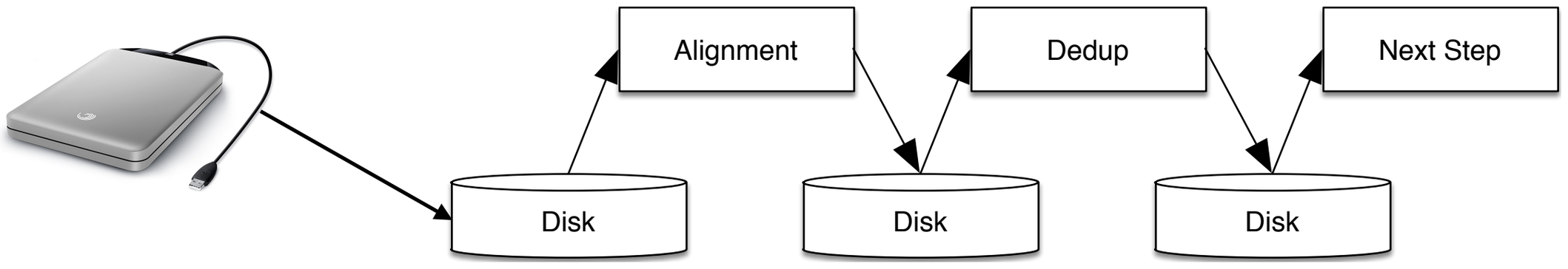
Analysis speed matters…



Saunders et. al., Sci Transl Med 2012

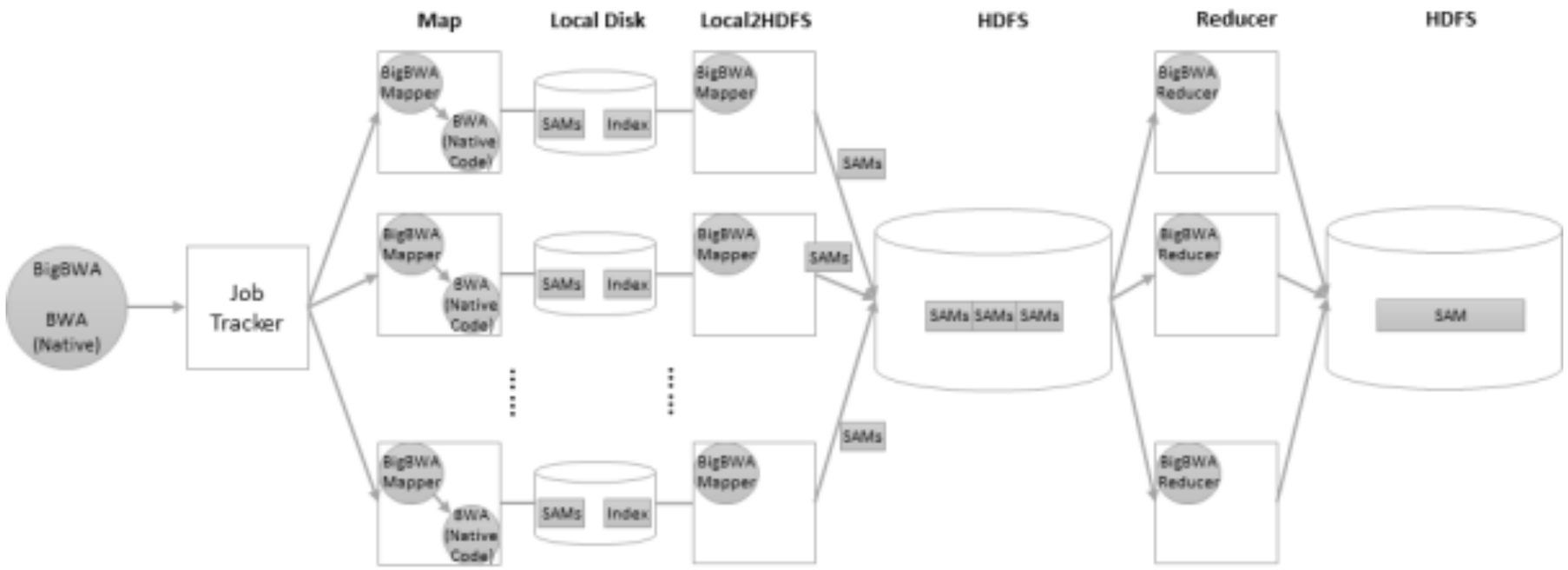
- Emergency genomics
- Tailormade drug using mutation pattern of patients
- The time for the analysis can be further improved

Sequence analysis pipeline



- TOO BIG DATA ~ 70 GB for a whole genome data
- The sequencing analysis pipeline is fragmented
    - Alignment(BWA) -> PCR duplication marking(Picard) -> Sort(Samtools) -> Base quality recalibration(GATK BaseRecalibrator) -> Variant calling(GATK HaplotypeCaller)
- Each step produces outputs and writes to the disk
    - Frequent time-consuming I/Os
- Each step is not parallelized enough for the best performance

Distributed computing based algorithms



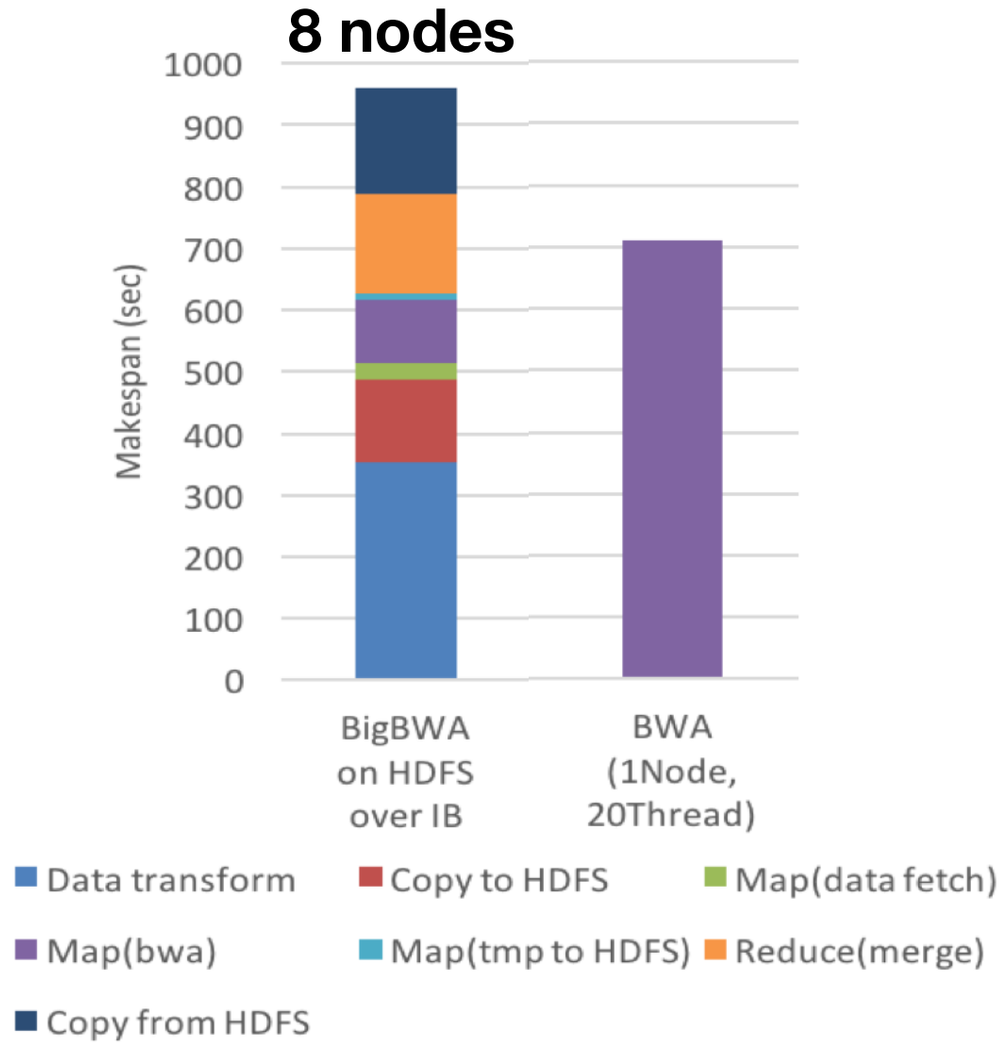
Abuín, J. M., et. al. (2015). Bioinformatics.

- Algorithms based on distributed computing frameworks(e.g. Hadoop, Spark, etc.)

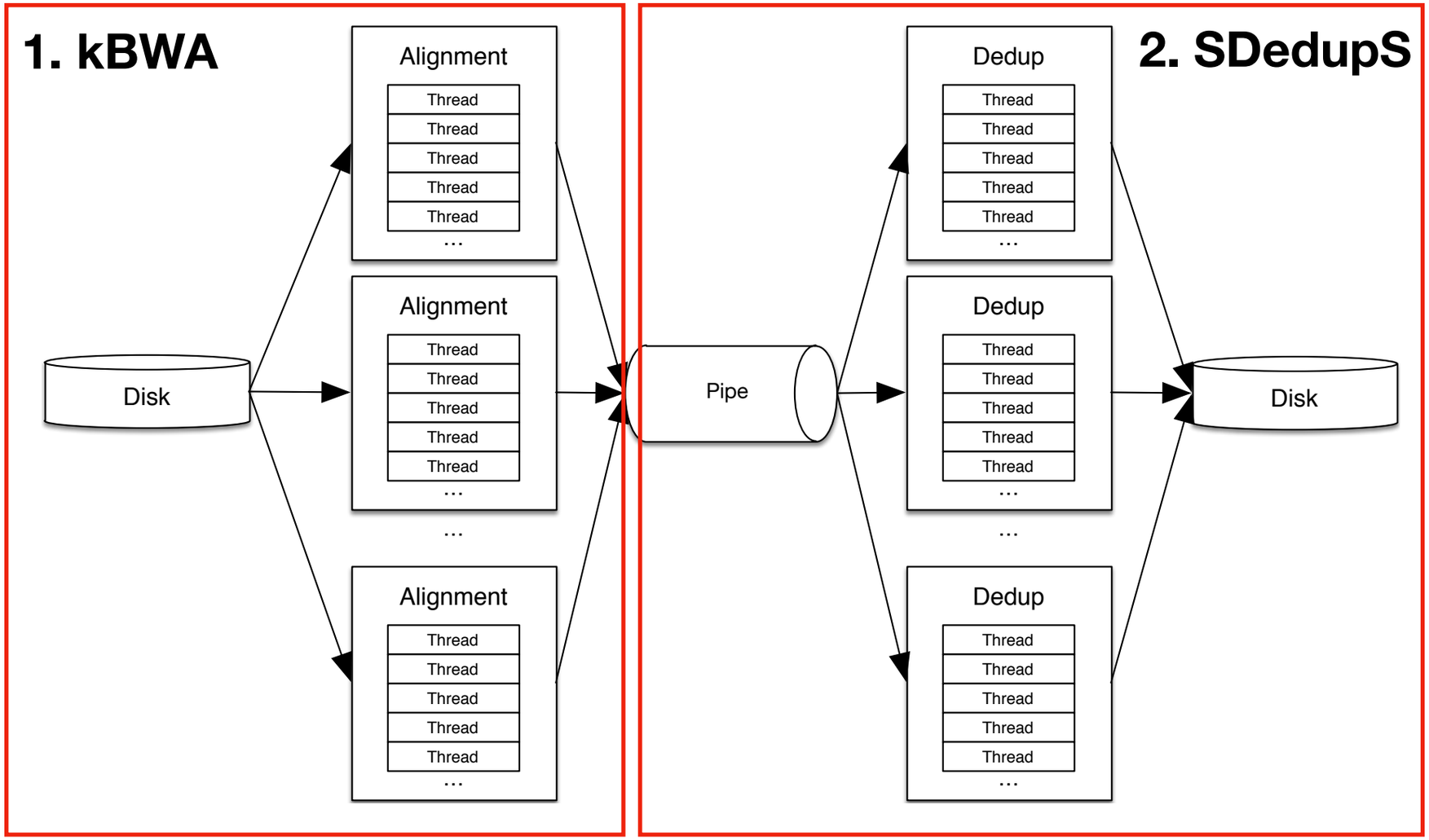
- Divides the job to sub-tasks and distributes to the multiple computing nodes for scalable parallelization

Pros and cons of MapReduce based algorithms

- Pros
  - Can be easily implemented in parallel processing by using predefined functions
  - Utilizes various job management functions of Hadoop for stable execution
- Cons
  - Unavoidable use of Hadoop File System(HDFS)
    - Overhead for transforming the data and splitting & merging data

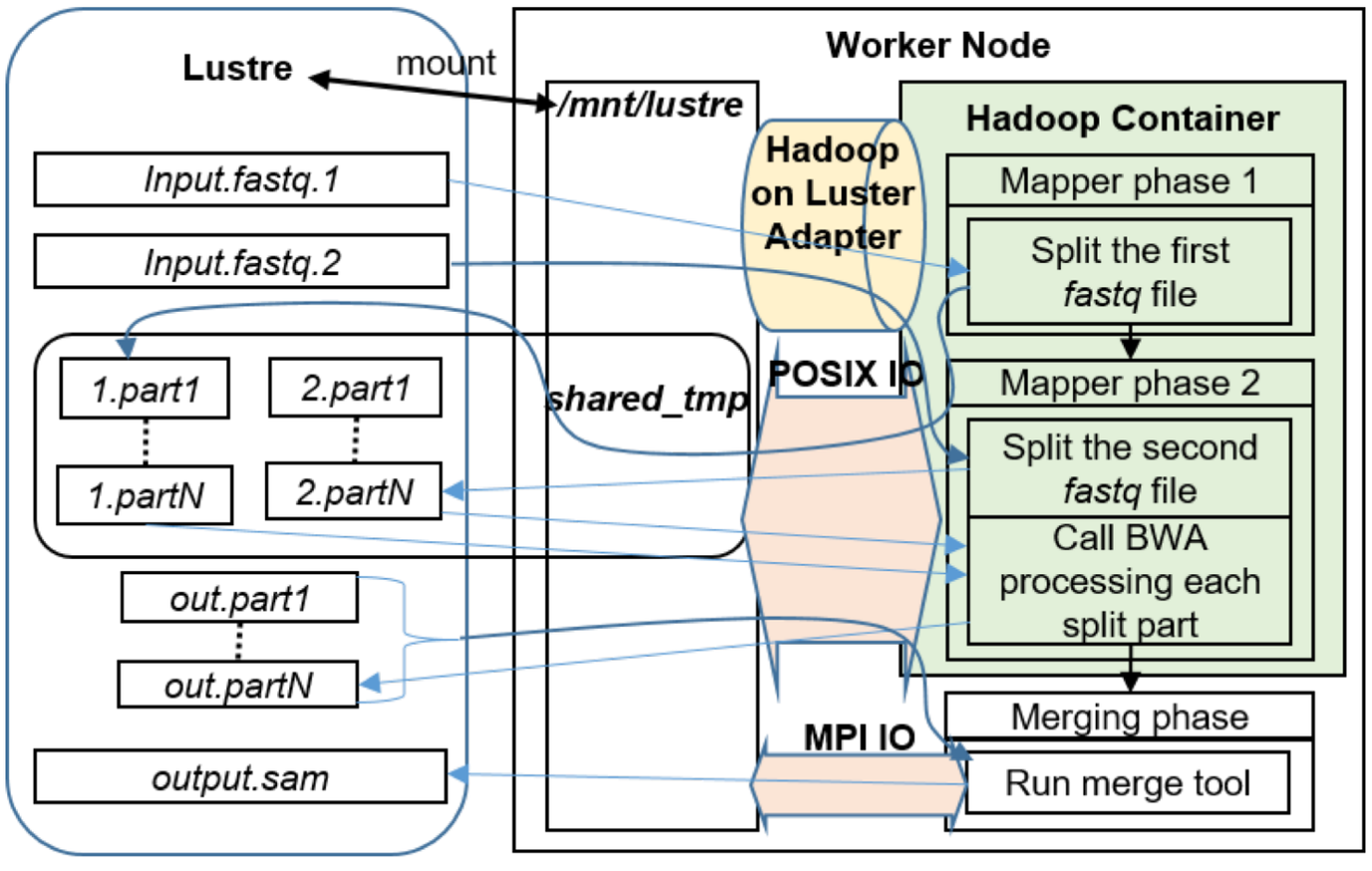


Parallel distributed sequence analysis pipeline with streaming



- Using the distributed computing frameworks for scalable and reliable parallelization
- Using popular algorithms as it is, as possible (BWA, Samblaster, etc)
- Reducing the I/O overhead for data transform and movement caused by HDFS

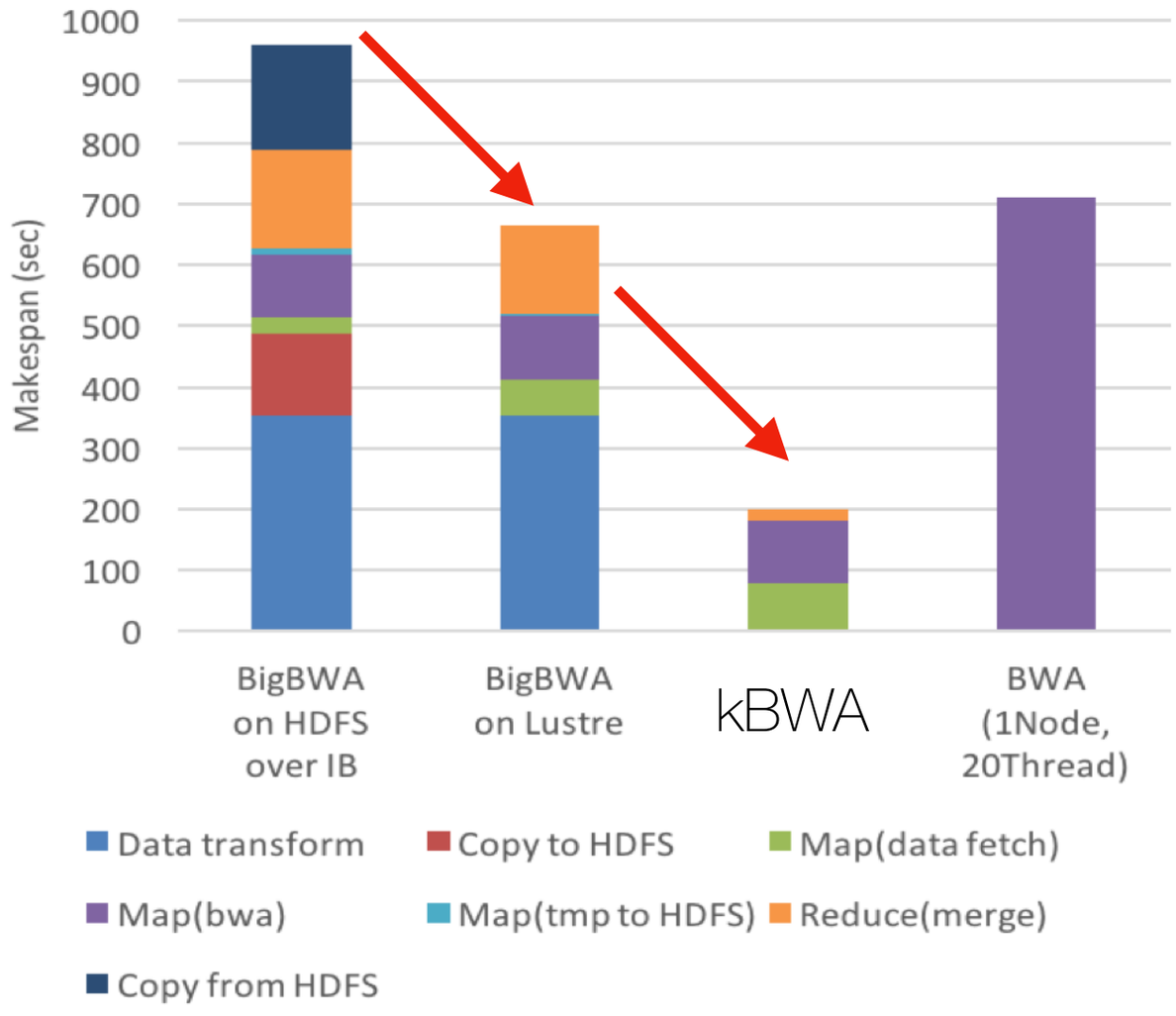
kBWA



- By modifying and enhancing BigBWA
- Used Lustre file system instead of HDFS by adapting Hadoop on Lustre plug-in
  - No need to use HDFS; no copy between your storage and HDFS are required
  - Can use Shared temp directory among the mappers
- Used MPI-IO to parallelize the reduce-phase that merges the aligned sam files

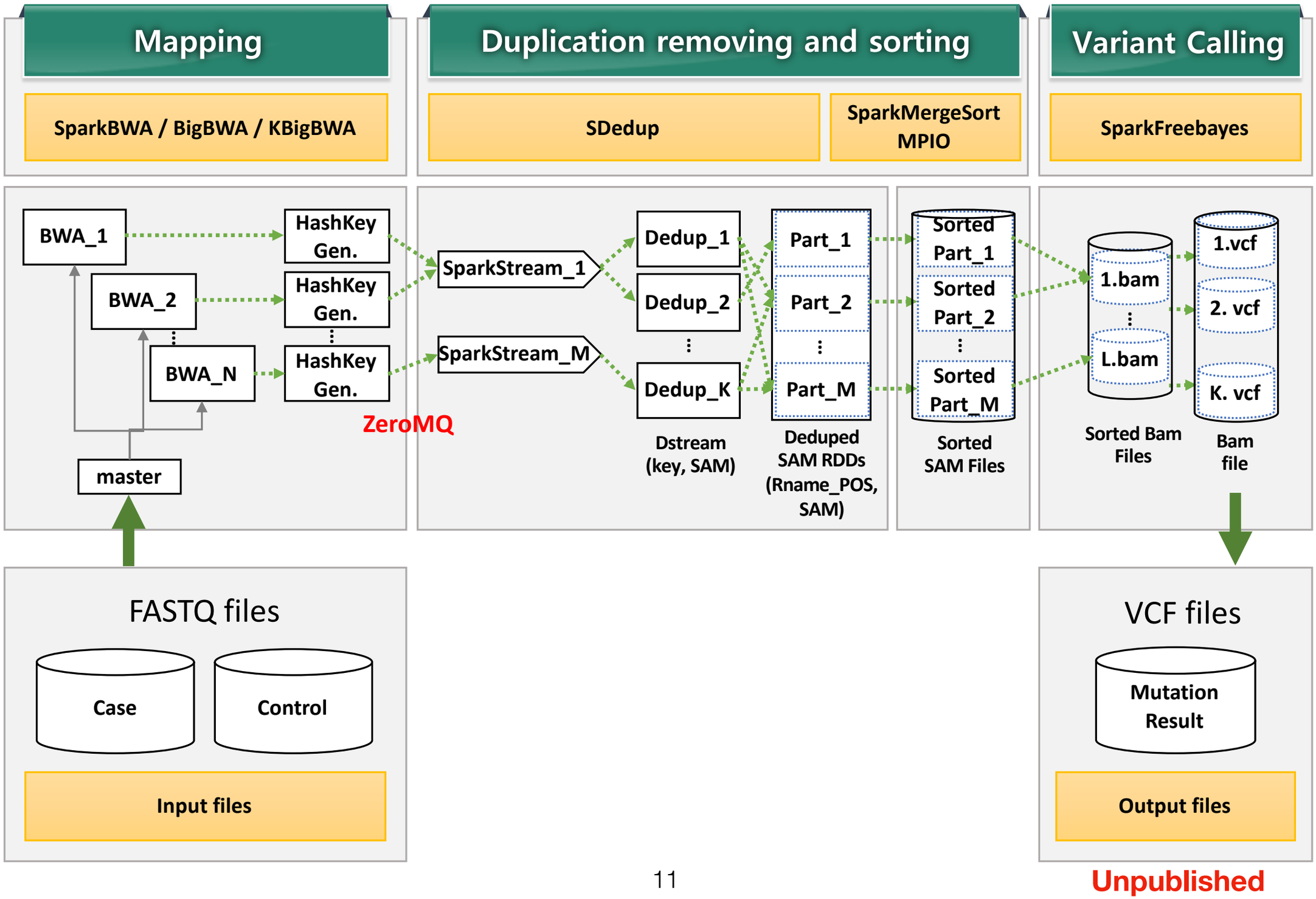
kBWA performance

- 8 nodes with Lustre file system
- 142% faster when using Hadoop on Lustre plug-in
- Eliminating data transform by using shared tmp directory and parallelizing reduce phase improve the performance further to 475%
  - 3.6 times faster than one node execution
  - Parallel efficiency: 0.45



Byun et. al., Accelerating Genome Sequence Alignment on Hadoop on Lustre Environment. in 436–437 (IEEE, 2017).

Spark-based parallel mutation calling pipeline



Unpublished

Results

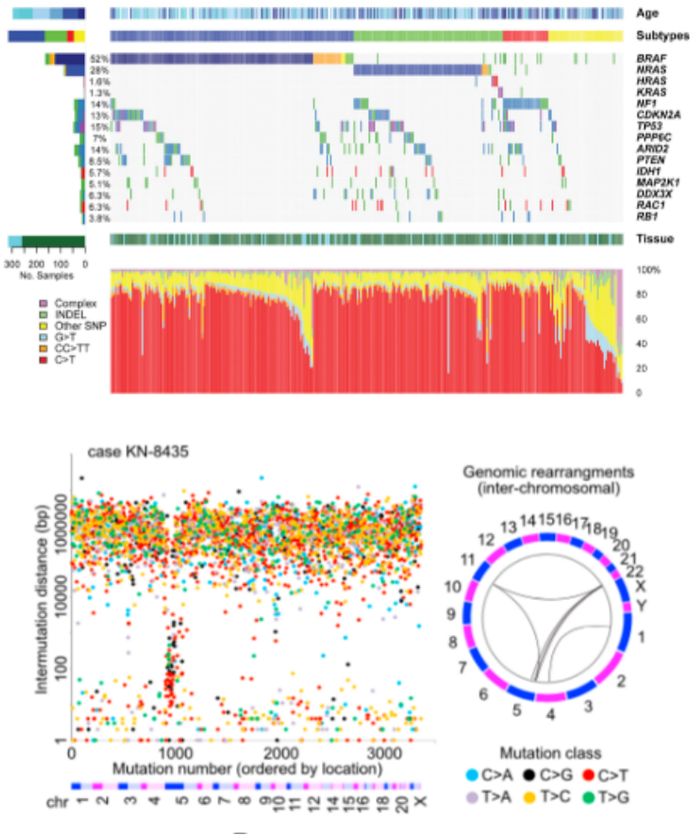
Resources		Specification
Cluster	H/W Spec	8 nodes, Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30GHz (10 cores), 80GiB
	S/W Spec	CentOS 7.3.16, Hadoop 2.7.3, Spark 2.3, ZeroMQ 4.3.1
Data	Input DNA sequence data	104 GB genome sequence raw data(FASTQ format, paired end)
	Referece genome data	3.0G hg19.fa

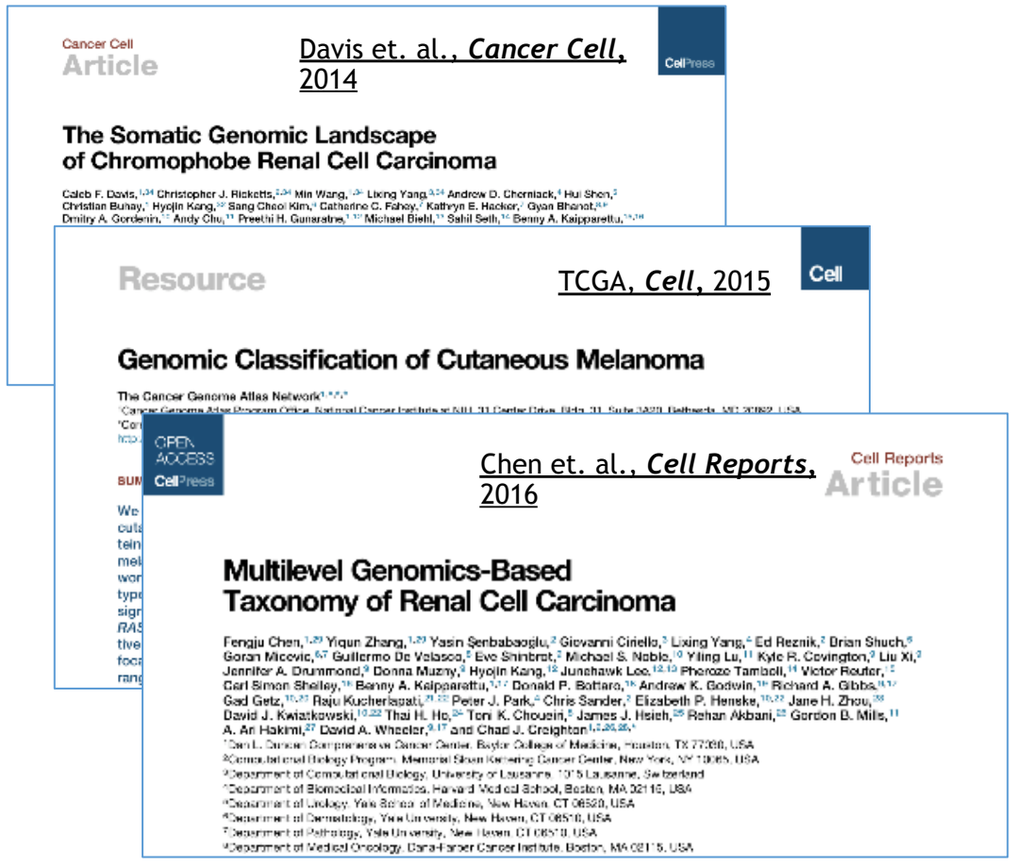
Nodes	Alignment	Dedup	Sort	Variant call	Total
1	48m19s	12m5s	31m42s	69m5s	**161m11s**
4	22m28s	15m30s	5m46s	15m11s	**58m31s**
8	12m25s	9m5s	3m36s	9m32s	**34m38s**

Unpublished data

- Our pipeline can complete the analysis pipeline in 35 minutes with 8 computing nodes
  - Achieved 467% speed up with 8 nodes (Parallel efficiency: 58.375)
- Currently under testing the developed pipeline with the variable data sizes and the number of nodes

Global cancer genome research collaborations




Cancer Cell
Article

Davis et. al., ***Cancer Cell***, 2014

The Somatic Genomic Landscape of Chromophobe Renal Cell Carcinoma

Resource

TCGA, ***Cell***, 2015

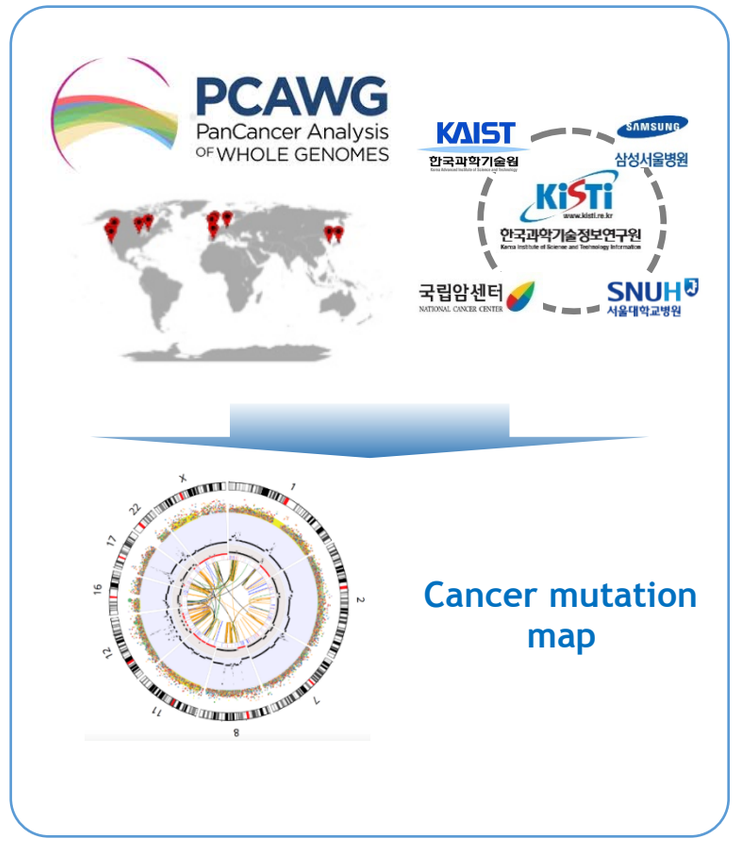
Genomic Classification of Cutaneous Melanoma

OPEN ACCESS

Cell Reports
Article

Chen et. al., ***Cell Reports***, 2016

Multilevel Genomics-Based Taxonomy of Renal Cell Carcinoma



- The Cancer Genome Atlas(TCGA) collaboration studies
  - ~ 500 samples for each cancer type (~ 120 TB for each cancer type)
  - Chromosome structure variation detection for the kidney cancer and the skin cancer projects
- Planned to start analyzing PanCancer Analysis of Whole Genome(PCAWG) project's data in 2019
  - ~ 6,000 samples for various types of cancers (~ 1.4PB of data)

Korean lung cancer studies

JOURNAL OF CLINICAL ONCOLOGY | BIOLOGY OF NEOPLASIA

Clonal History and Genetic Predictors of Transformation Into Small-Cell Carcinomas From Lung Adenocarcinomas

June-Koo Lee, Junehawk Lee, Sehui Kim, Soyeon Kim, Jeonghwan Youk, Seongyeol Park, Yohan An, Bhumsuk Keam, Dong-Wan Kim, Dae Seog Heo, Young Tae Kim, Jin-Soo Kim, Se Hyun Kim, Jong Seok Lee, Se-Hoon Lee, Keunchil Park, Ja-Lok Ku, Yoon Kyung Jeon, Doo Hyun Chung, Peter J. Park, Joon Kim, Tae Min Kim, and Young Seok Ju

Author affiliations and support information (if applicable) appear at the end of this article.

Published at jco.org on May 12, 2017.

T.M.K. and Y.S.J. are principal investigators who contributed equally to this study.

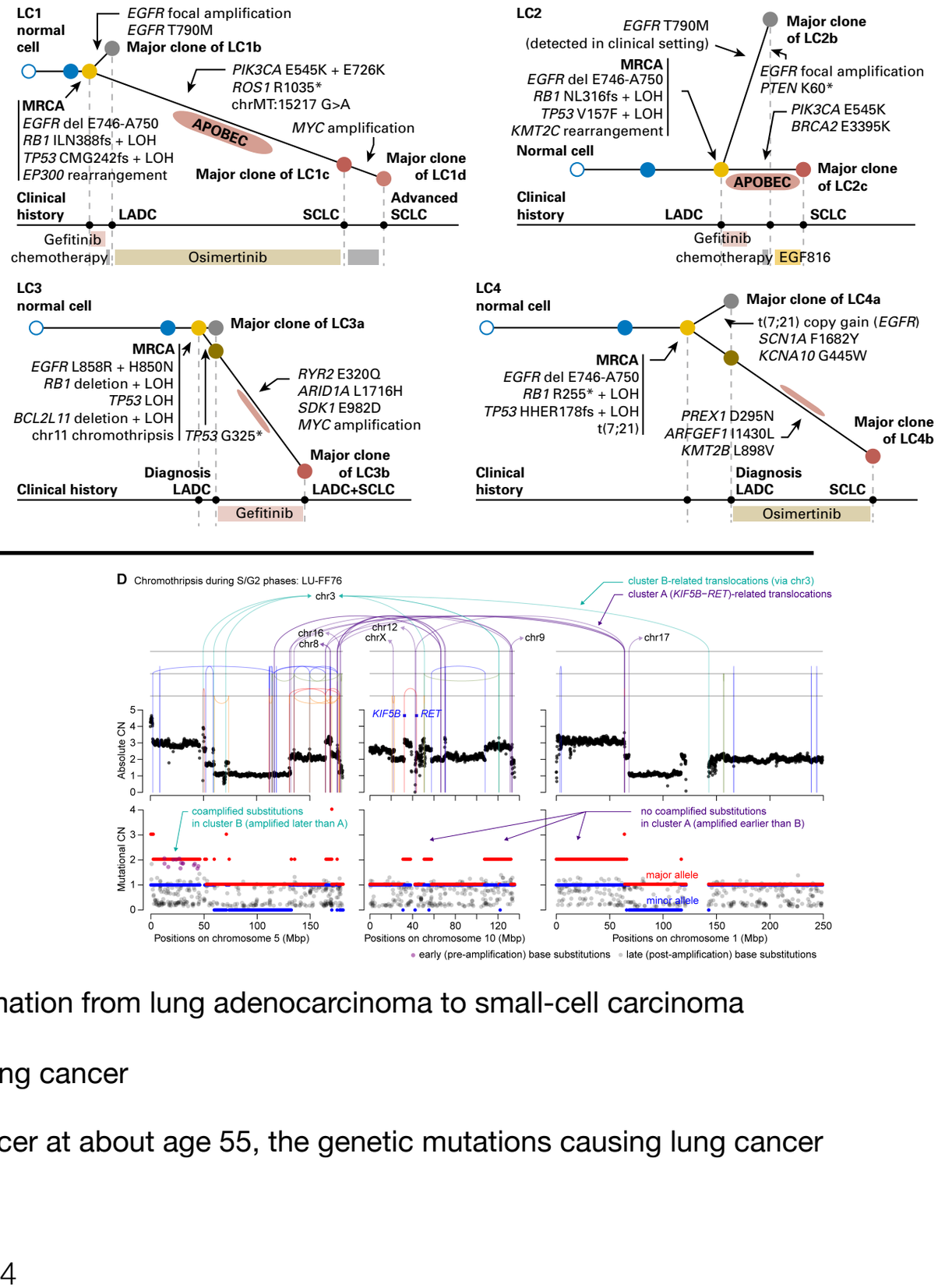
Corresponding author: Tae Min Kim, MD,

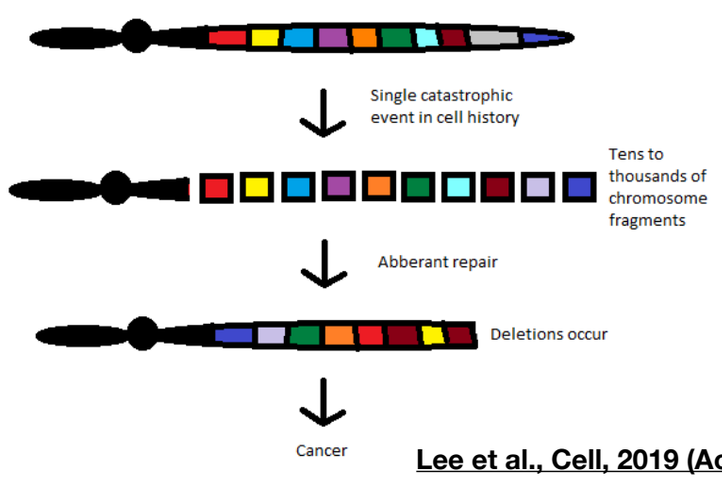
A B S T R A C T

Purpose

Histologic transformation of *EGFR* mutant lung adenocarcinoma (LADC) into small-cell lung cancer (SCLC) has been described as one of the major resistant mechanisms for epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors (TKIs). However, the molecular pathogenesis is still unclear.

Lee et. al., Journal of Clinical Oncology, 2017





Lee et al., Cell, 2019 (Accepted)

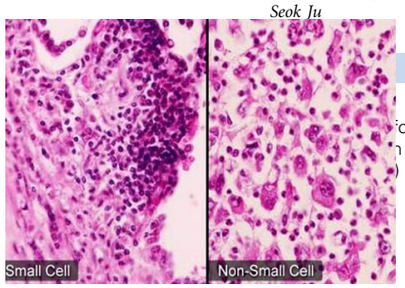
- Identification of genetic mutations highly correlated to transformation from lung adenocarcinoma to small-cell carcinoma
- Identification of complex rearrangements of chromosomes in lung cancer
    - Observed that while patients are diagnosed as lung cancer at about age 55, the genetic mutations causing lung cancer usually occur around age 27

Korean lung cancer studies

JOURNAL OF CLINICAL ONCOLOGY | BIOLOGY OF NEOPLASIA

Clonal History and Genetic Predictors of Transformation Into Small-Cell Carcinomas From Lung Adenocarcinomas

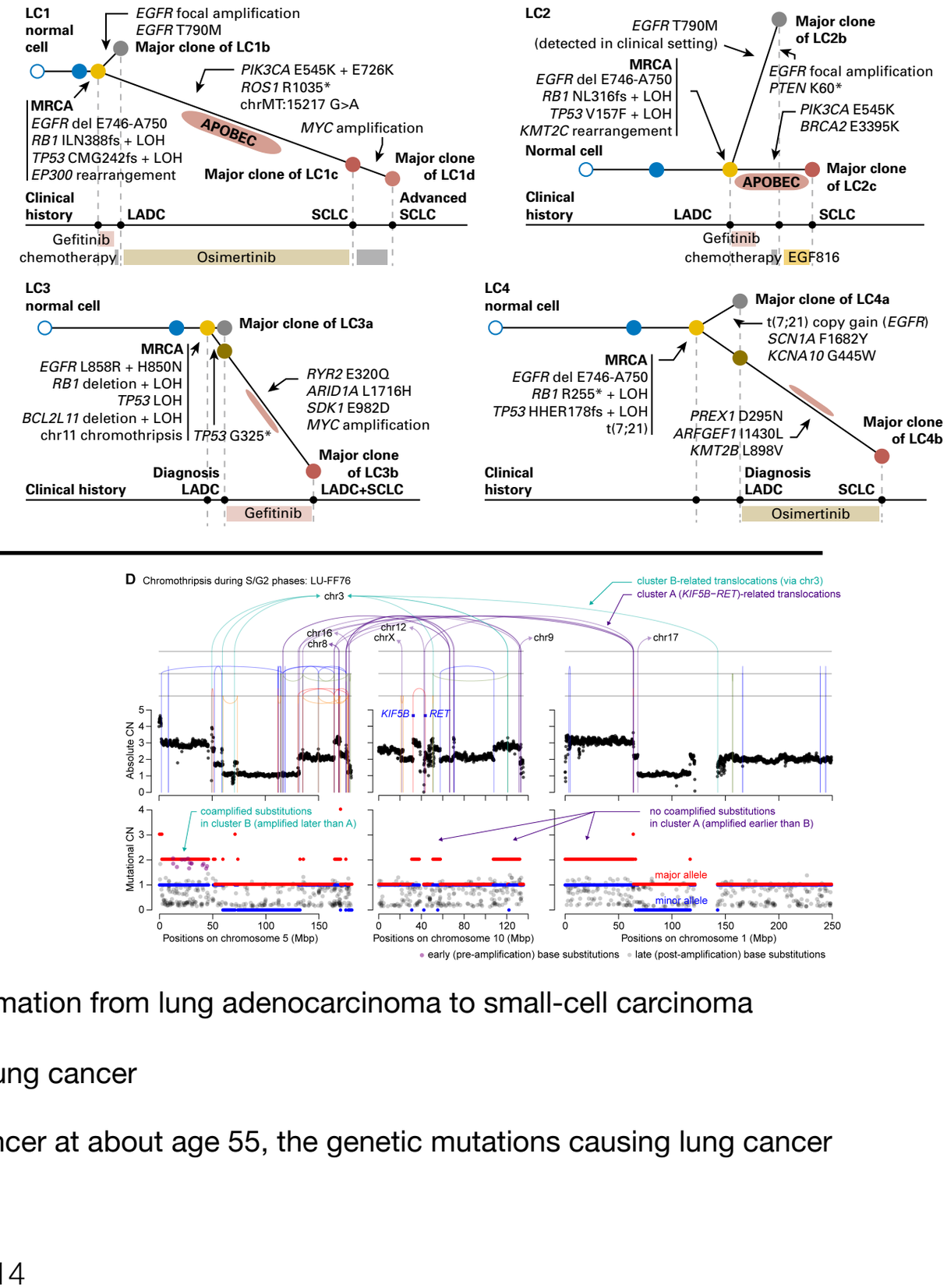
June-Koo Lee, Junehawk Lee, Sehui Kim, Soyeon Kim, Jeonghwan Youk, Seongyeol Park, Yohan An, Bhumsuk Keam, Dong-Wan Kim, Dae Seog Heo, Young Tae Kim, Jin-Soo Kim, Se Hyun Kim, Jong Seok Lee, Se-Hoon Lee, Keunchil Park, Ja-Lok Ku, Yoon Kyung Jeon, Doo Hyun Chung, Peter J. Park, Joon Kim, Tae Min Kim, and Young Seok Ju

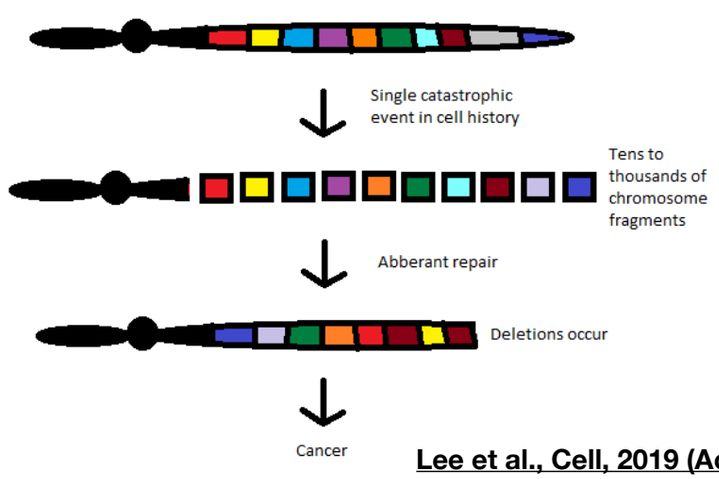


A B S T R A C T

ormation of *EGFR* mutant lung adenocarcinoma (LADC) into small-cell lung cancer described as one of the major resistant mechanisms for epidermal growth factor tyrosine kinase inhibitors (TKIs). However, the molecular pathogenesis is still

Lee et. al., Journal of Clinical Oncology, 2017





Lee et al., Cell, 2019 (Accepted)

- Identification of genetic mutations highly correlated to transformation from lung adenocarcinoma to small-cell carcinoma
- Identification of complex rearrangements of chromosomes in lung cancer
    - Observed that while patients are diagnosed as lung cancer at about age 55, the genetic mutations causing lung cancer usually occur around age 27

Korean lung cancer studies

JOURNAL OF CLINICAL ONCOLOGY | BIOLOGY OF NEOPLASIA

Clonal History and Genetic Predictors of Transformation Into Small-Cell Carcinomas From Lung Adenocarcinomas

June-Koo Lee, Junehawk Lee, Sehui Kim, Soyeon Kim, Jeonghwan Youk, Seongyeol Park, Yohan An, Bhumsuk Keam, Dong-Wan Kim, Dae Seog Heo, Young Tae Kim, Jin-Soo Kim, Se Hyun Kim, Jong Seok Lee, Se-Hoon Lee, Keunchil Park, Ja-Lok Ku, Yoon Kyung Jeon, Doo Hyun Chung, Peter J. Park, Joon Kim, Tae Min Kim, and Young Seok Ju

Author affiliations and support information (if applicable) appear at the end of this article.

Published at jco.org on May 12, 2017.

T.M.K. and Y.S.J. are principal investigators who contributed equally to this study.

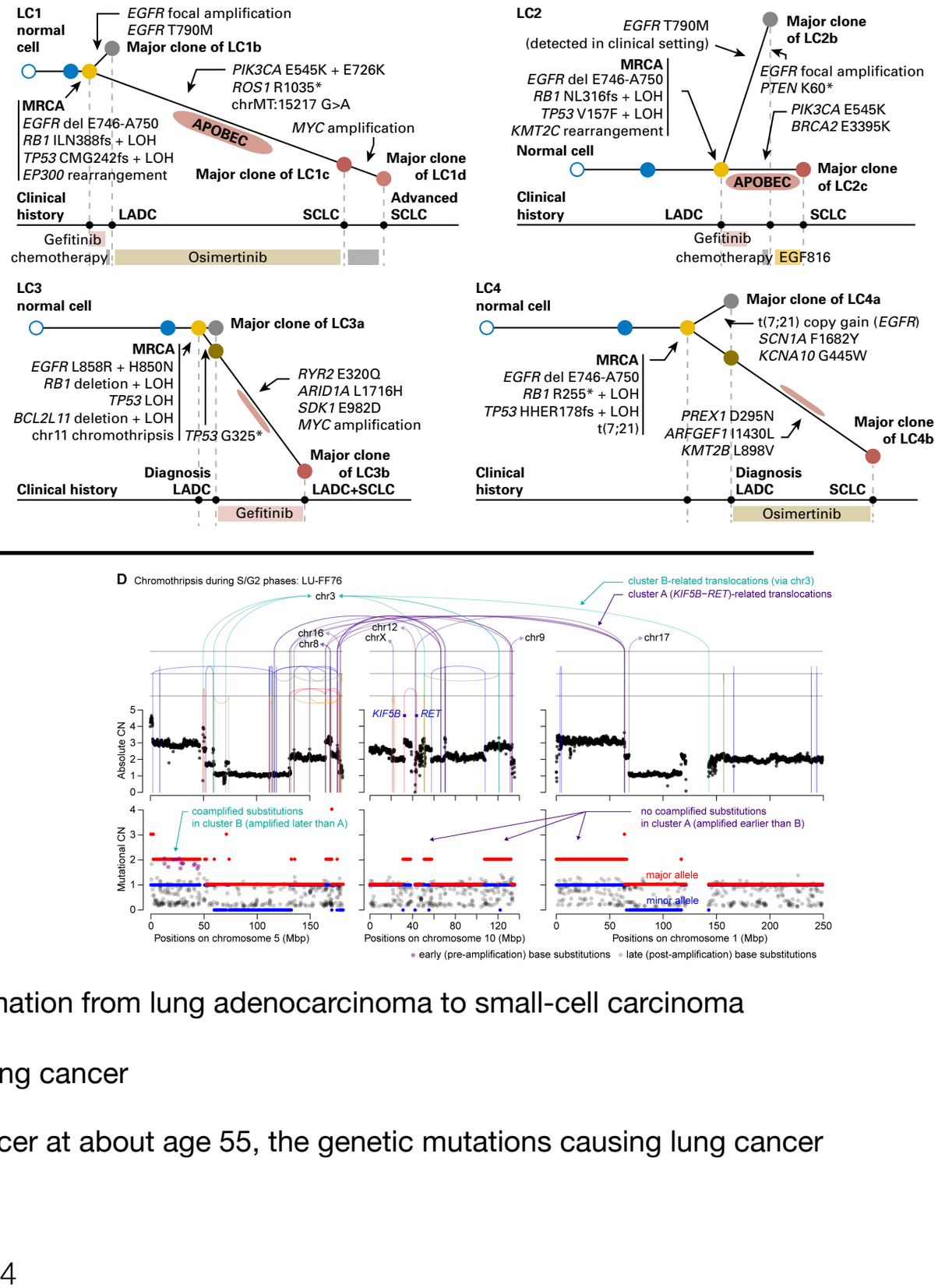
Corresponding author: Tae Min Kim, MD,

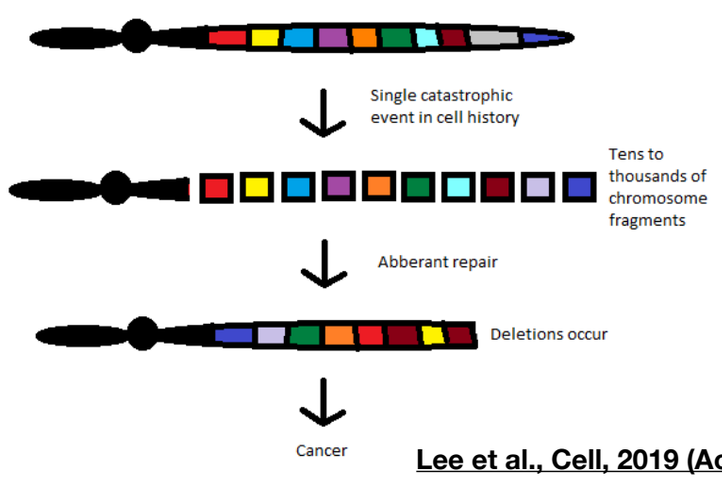
A B S T R A C T

Purpose

Histologic transformation of *EGFR* mutant lung adenocarcinoma (LADC) into small-cell lung cancer (SCLC) has been described as one of the major resistant mechanisms for epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors (TKIs). However, the molecular pathogenesis is still unclear.

Lee et. al., Journal of Clinical Oncology, 2017





Lee et al., Cell, 2019 (Accepted)

- Identification of genetic mutations highly correlated to transformation from lung adenocarcinoma to small-cell carcinoma
- Identification of complex rearrangements of chromosomes in lung cancer
    - Observed that while patients are diagnosed as lung cancer at about age 55, the genetic mutations causing lung cancer usually occur around age 27

Korean lung cancer studies

JOURNAL OF CLINICAL ONCOLOGY | BIOLOGY OF NEOPLASIA

Clonal History and Genetic Predictors of Transformation Into Small-Cell Carcinomas From Lung Adenocarcinomas

June-Koo Lee, Junehawk Lee, Sehui Kim, Soyeon Kim, Jeonghwan Youk, Seongyeol Park, Yohan An, Bhumsuk Keam, Dong-Wan Kim, Dae Seog Heo, Young Tae Kim, Jin-Soo Kim, Se Hyun Kim, Jong Seok Lee, Se-Hoon Lee, Keunchil Park, Ja-Lok Ku, Yoon Kyung Jeon, Doo Hyun Chung, Peter J. Park, Joon Kim, Tae Min Kim, and Young Seok Ju

Author affiliations and support information (if applicable) appear at the end of this article.

Published at jco.org on May 12, 2017.

T.M.K. and Y.S.J. are principal investigators who contributed equally to this study.

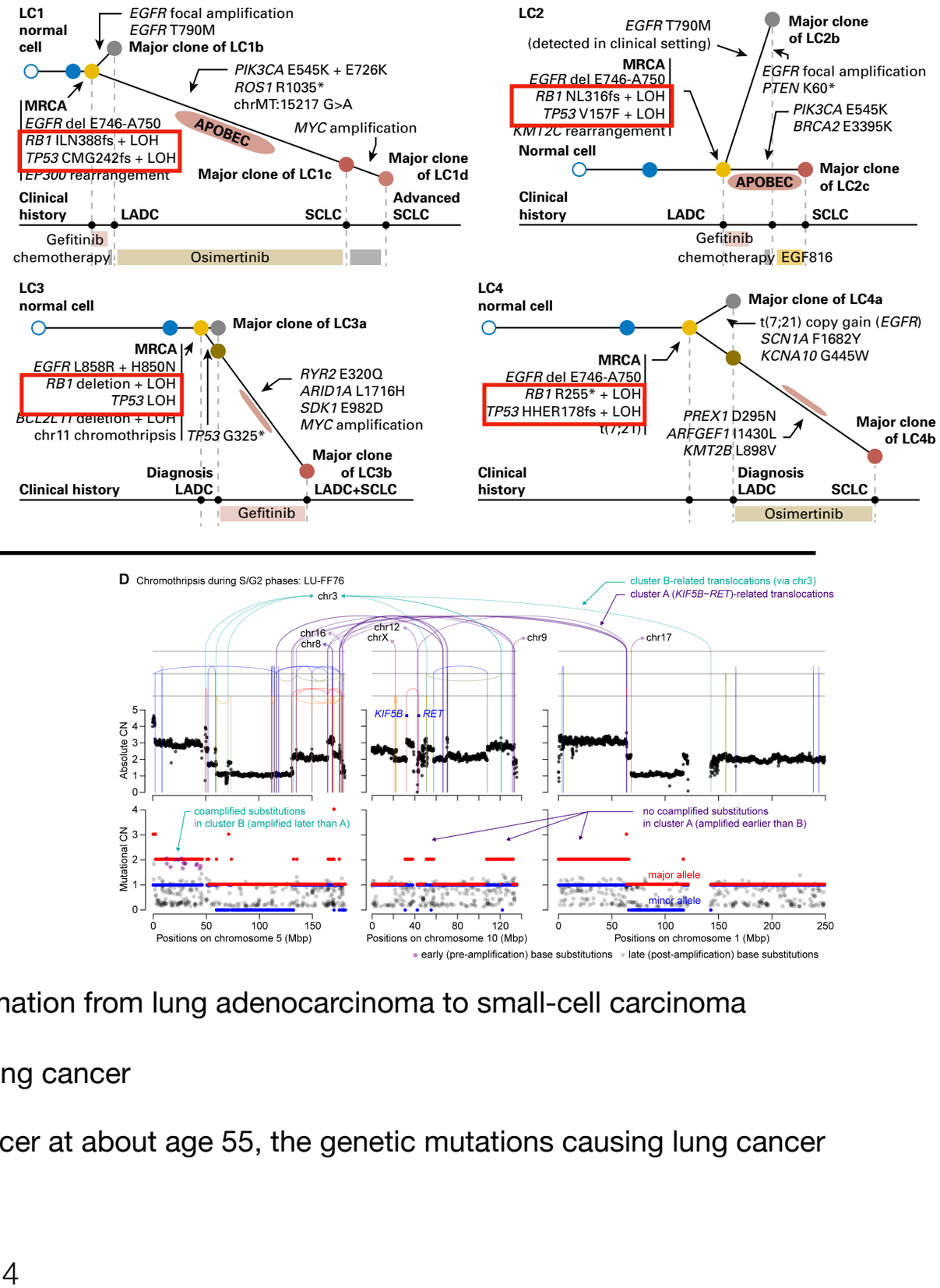
Corresponding author: Tae Min Kim, MD,

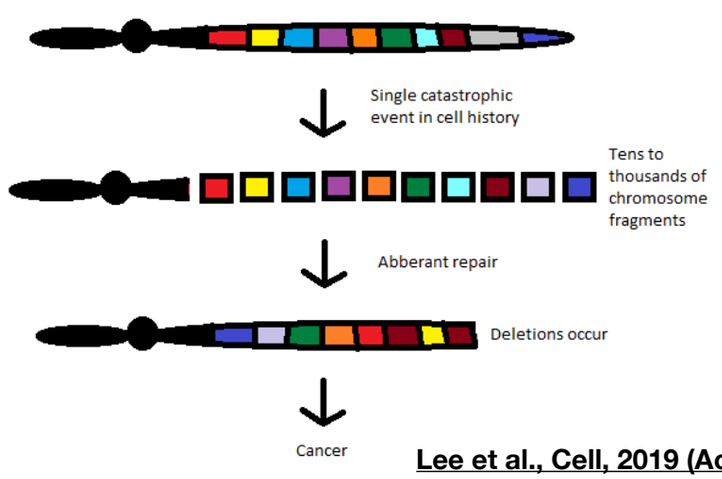
A B S T R A C T

Purpose

Histologic transformation of *EGFR* mutant lung adenocarcinoma (LADC) into small-cell lung cancer (SCLC) has been described as one of the major resistant mechanisms for epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors (TKIs). However, the molecular pathogenesis is still unclear.

Lee et. al., Journal of Clinical Oncology, 2017





Lee et al., Cell, 2019 (Accepted)

- Identification of genetic mutations highly correlated to transformation from lung adenocarcinoma to small-cell carcinoma
- Identification of complex rearrangements of chromosomes in lung cancer
    - Observed that while patients are diagnosed as lung cancer at about age 55, the genetic mutations causing lung cancer usually occur around age 27

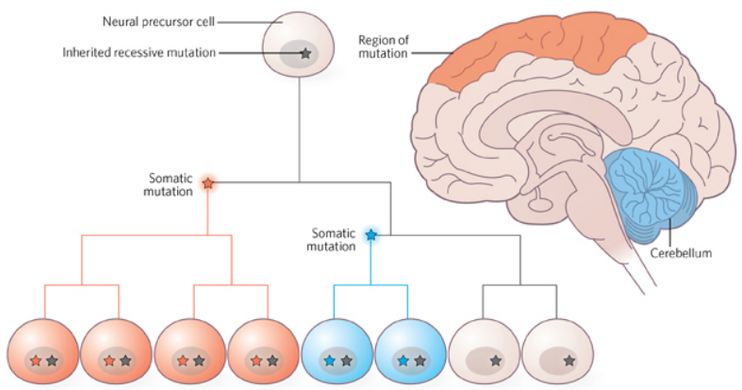
Somatic mutations in brain

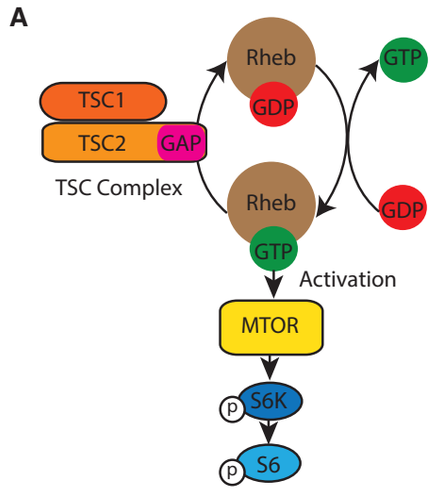
ARTICLE

Somatic Mutations in *TSC1* and *TSC2* Cause Focal Cortical Dysplasia

Jae Seok Lim,[1,13] Ramu Gopalappa,[2,3,13] Se Hoon Kim,[4,13] Suresh Ramakrishna,[2] Minji Lee,[5] Woo-il Kim,[1] Junho Kim,[7] Sang Min Park,[1] Junehawk Lee,[8] Jung-Hwa Oh,[9] Heung Dong Kim,[10] Chang-Hwan Park,[2] Joon Soo Lee,[10] Sangwoo Kim,[7] Dong Seok Kim,[11] Jung Min Han,[5,6] Hoon-Chul Kang,[10,14] Hyongbum (Henry) Kim,[3,7,12,14,*] and Jeong Ho Lee[1,14,*]

Focal cortical dysplasia (FCD) is a major cause of the sporadic form of intractable focal epilepsies that require surgical treatment recently been reported that brain somatic mutations in *MTOR* account for 15%–25% of FCD type II (FCDII), characterized by dyslamination and dysmorphic neurons. However, the genetic etiologies of FCDII-affected individuals who lack the *MTOR* mu remain unclear. Here, we performed deep hybrid capture and amplicon sequencing (read depth of 100×–20,012×) of five imp mTOR pathway genes—*PIK3CA, PIK3R2, AKT3, TSC1,* and *TSC2*—by using paired brain and saliva samples from 40 FCDII indi negative for *MTOR* mutations. We found that 5 of 40 individuals (12.5%) had brain somatic mutations in *TSC1* (c.64C>T [p.Arg and c.610C>T [p.Arg204Cys]) and *TSC2* (c.4639G>A [p.Val1547Ile]), and these results were reproducible on two different sequencing

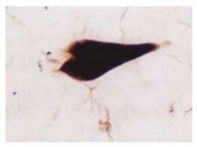




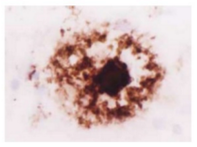
https://www.sciencedirect.com/science/article/pii/S0896627318304379?via%3Dihub

Lim et. al., Am. J. Hum. Genet., 2017

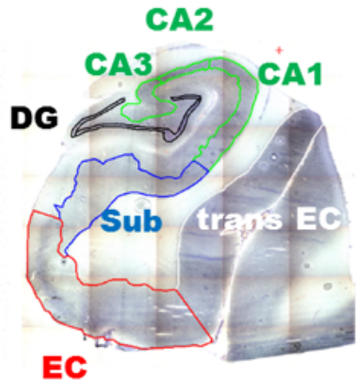
Park et. al., Neuron, 2018

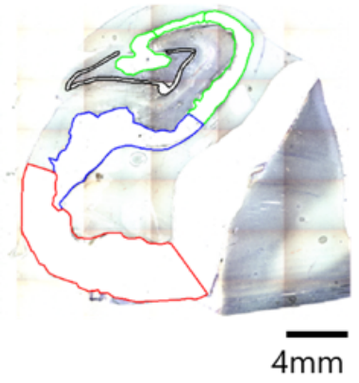


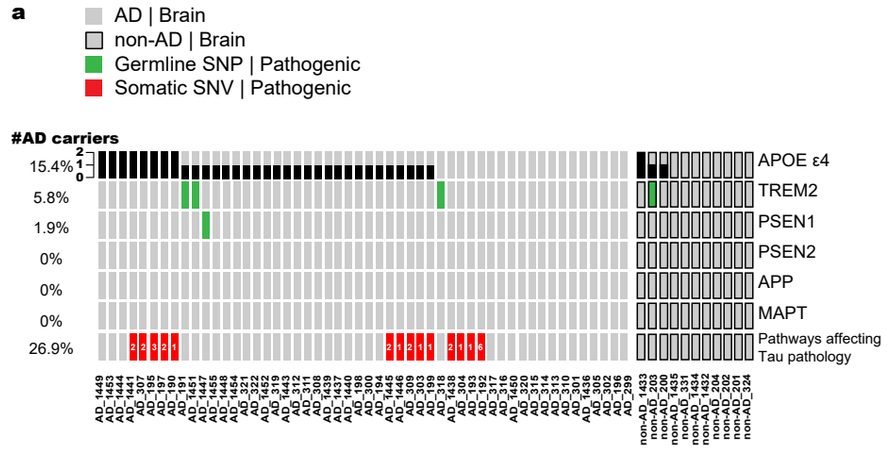
Neurofibrillary Tangles



Senile Plaques (β-amyloid)







Park and Lee et. al., (in revision)

- Identified somatic mutations in brain that can cause epilepsy
  - Found that the drugs targeting the mutated proteins are effective to treat epileptic seizure
- Identified somatic mutations in the brain of Alzheimer's disease patients

Summary

- Analyzing the deluge of genome sequence data requires speed and scale

- Parallel distributed framework based sequence analysis tools proposed to date, have limitations like restricted streamed executions and overheads from using HDFS

- By replacing HDFS with Lustre and by incorporating parallel distributed framework Spark, we accelerated genome sequence data analysis pipeline 4.7 times faster when using 8 computing nodes.

- KISTI is collaborating with global and domestic genomics researchers by providing supercomputing resources along with fast and scalable genome analysis pipelines

KiSTi
www.kisti.re.kr
한국과학기술정보연구원
Korea Institute of Science and Technology Information

Thank you