HUAWEI

The Hong Kong Polytechnic University
香港理工大學

LEADING NEW ICT

Next Generation Sequencing
&
Bio/Medical Data Analysis
Best Practice Sharing

Agenda

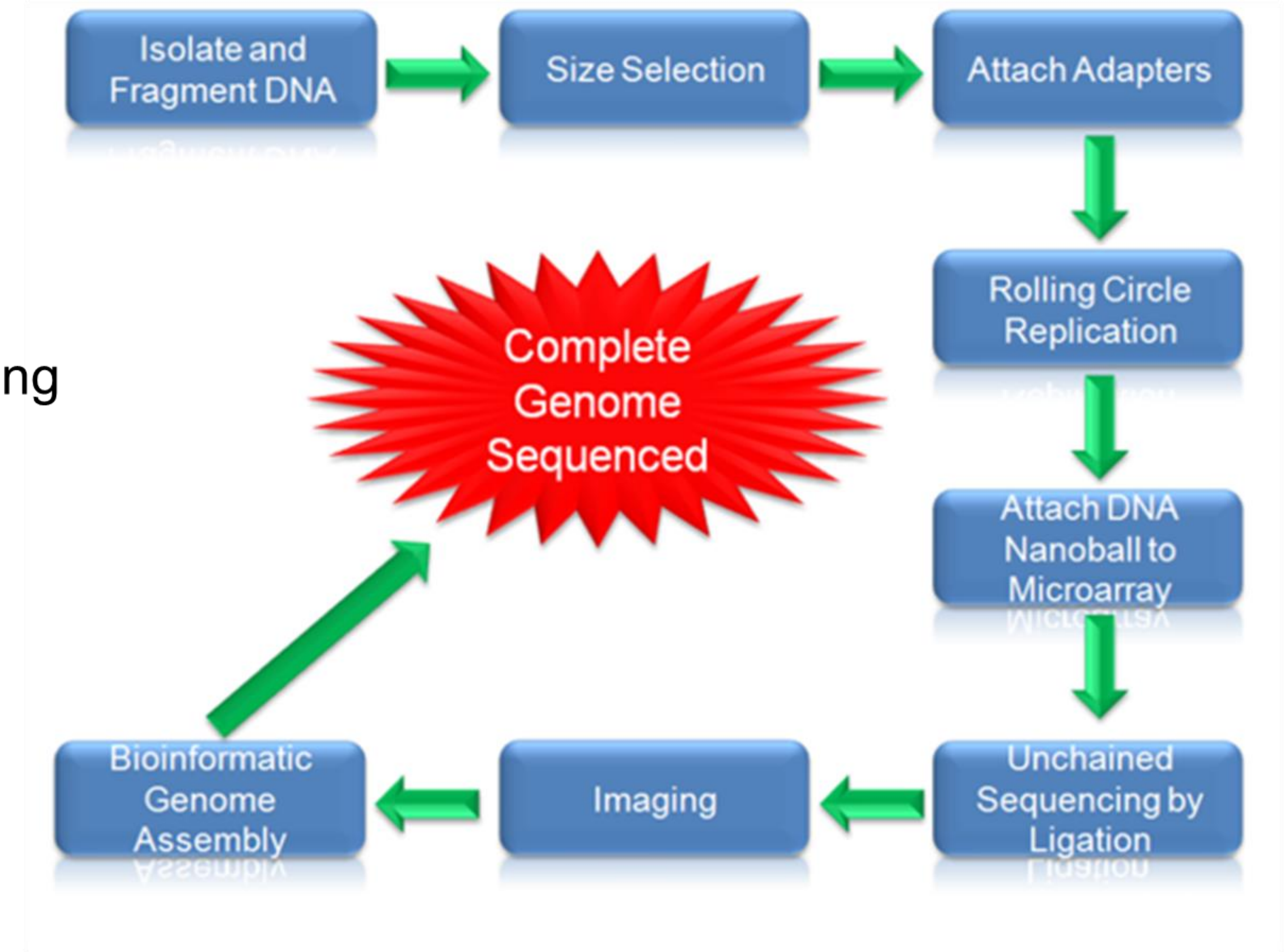
Part I – Presented by Dr. Cesar, PolyU

- Advanced Technology Platform:Next Generation Sequencing

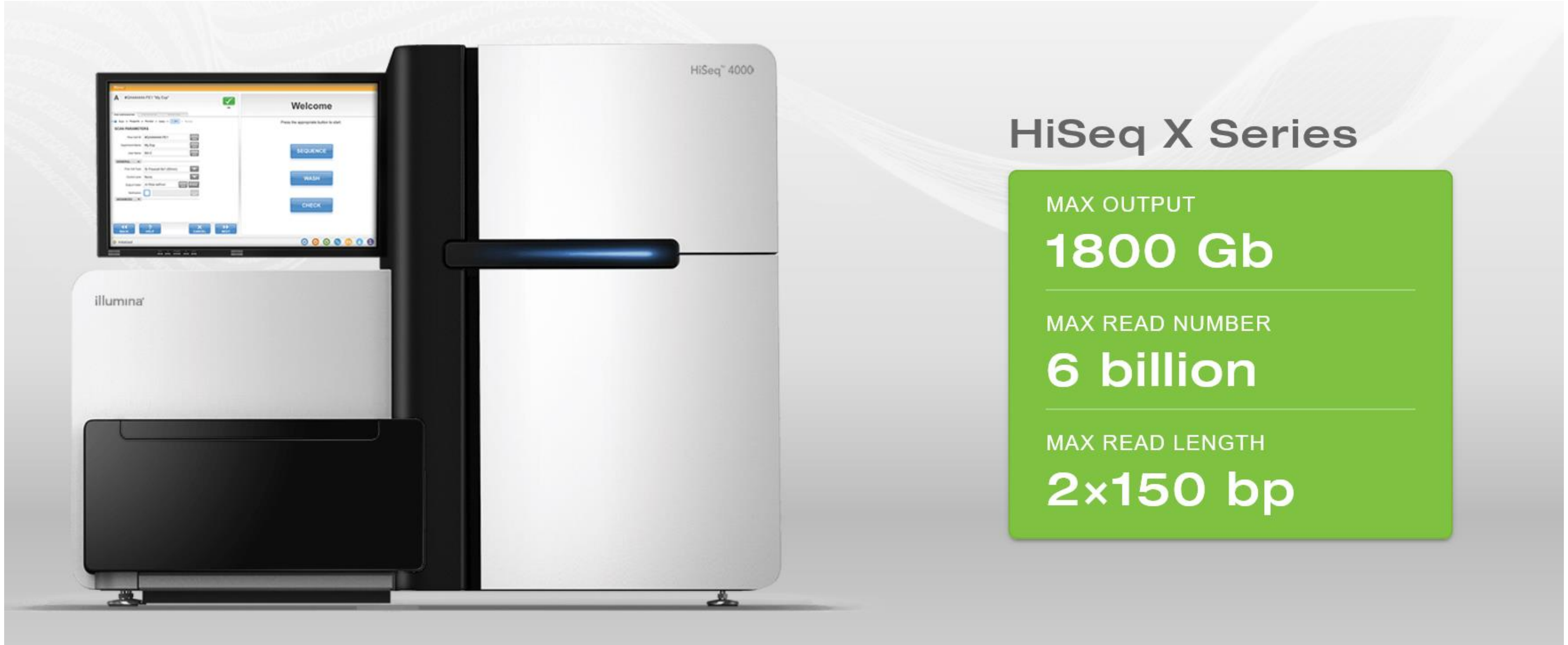
Part II – Presented by Yan Jun, Huawei

- Huawei Big Data Introduction
- Bio Information/Medical Big Data Practices

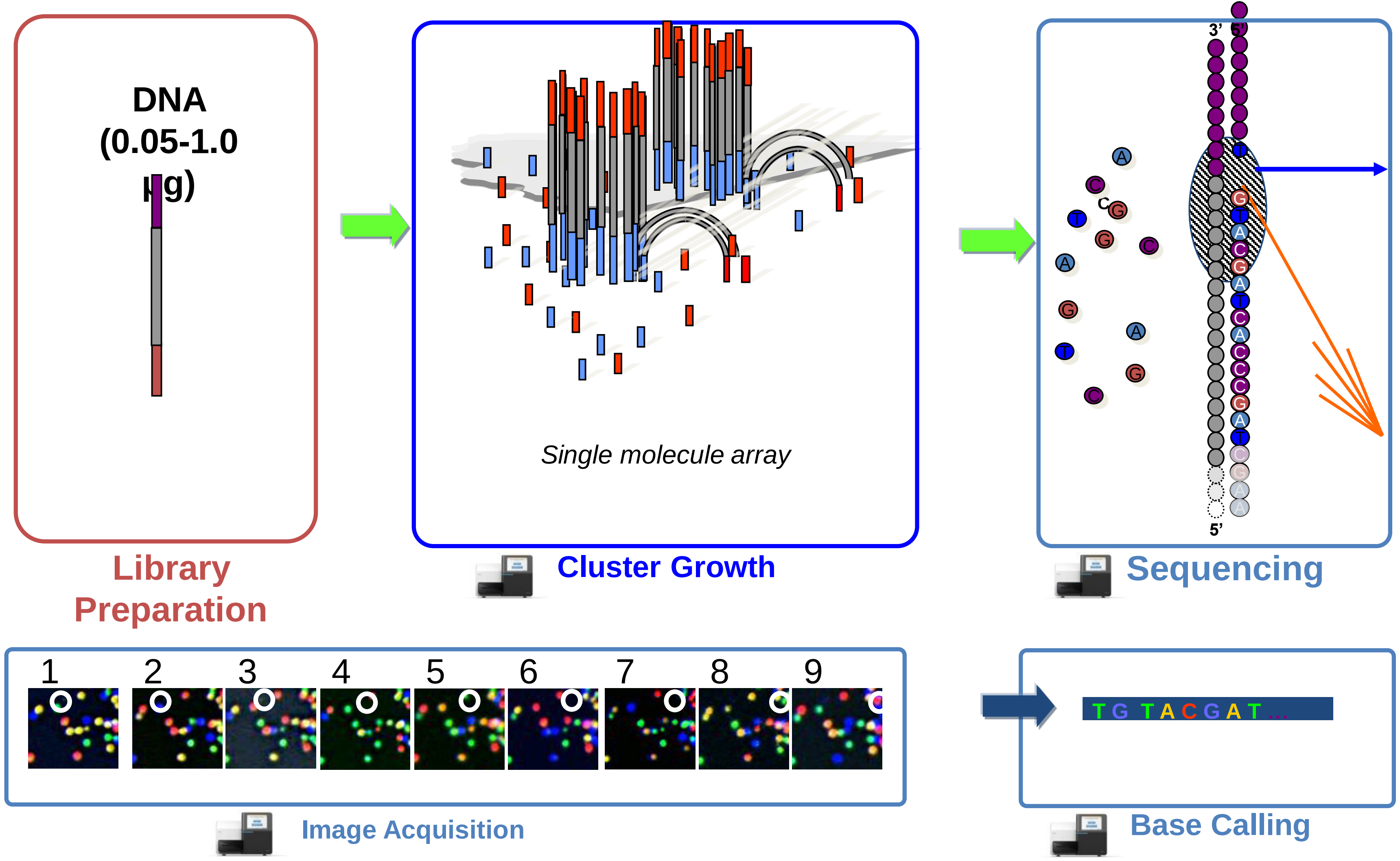
Next Generation Sequencing Procedures


Isolate and Fragment DNA
Size Selection
Attach Adapters
Rolling Circle Replication
Attach DNA Nanoball to Microarray
Unchained Sequencing by Ligation
Imaging
Bioinformatic Genome Assembly
Complete Genome Sequenced

Next Generation Sequencer



Illumine Sequencing Workflow



DNA (0.05-1.0 μg)

Single molecule array

Library Preparation

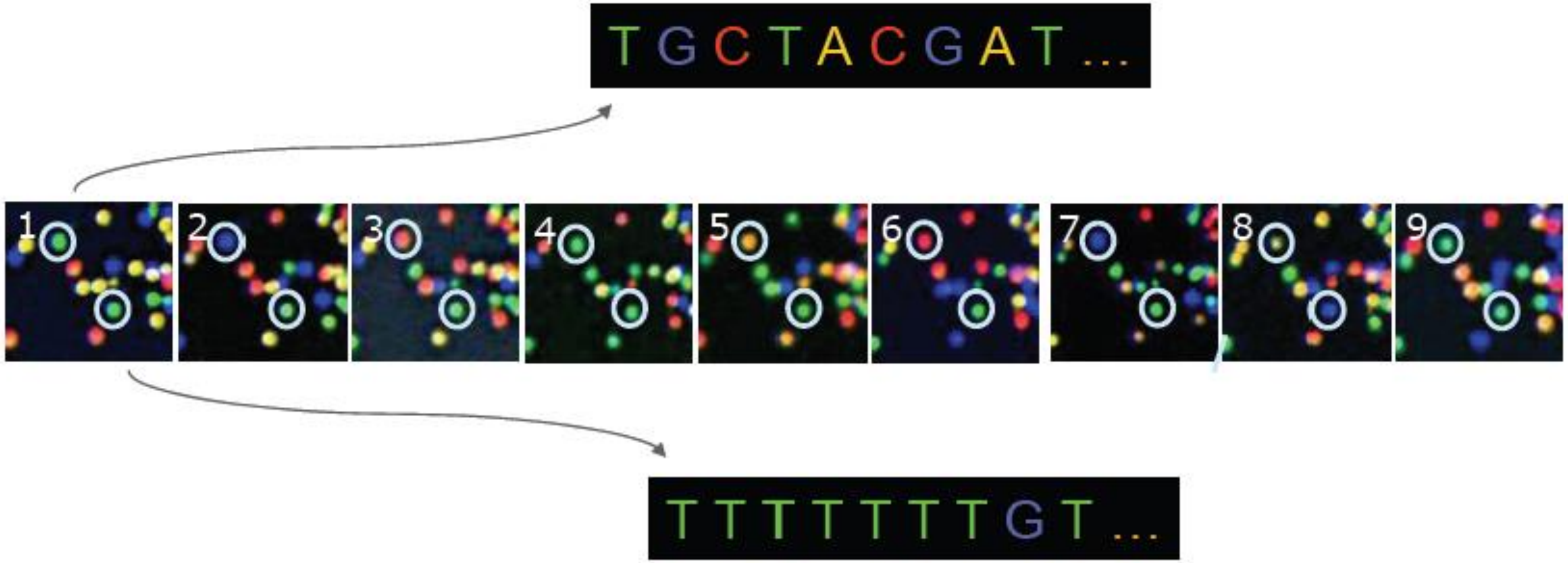
Cluster Growth

Sequencing

Image Acquisition

Base Calling

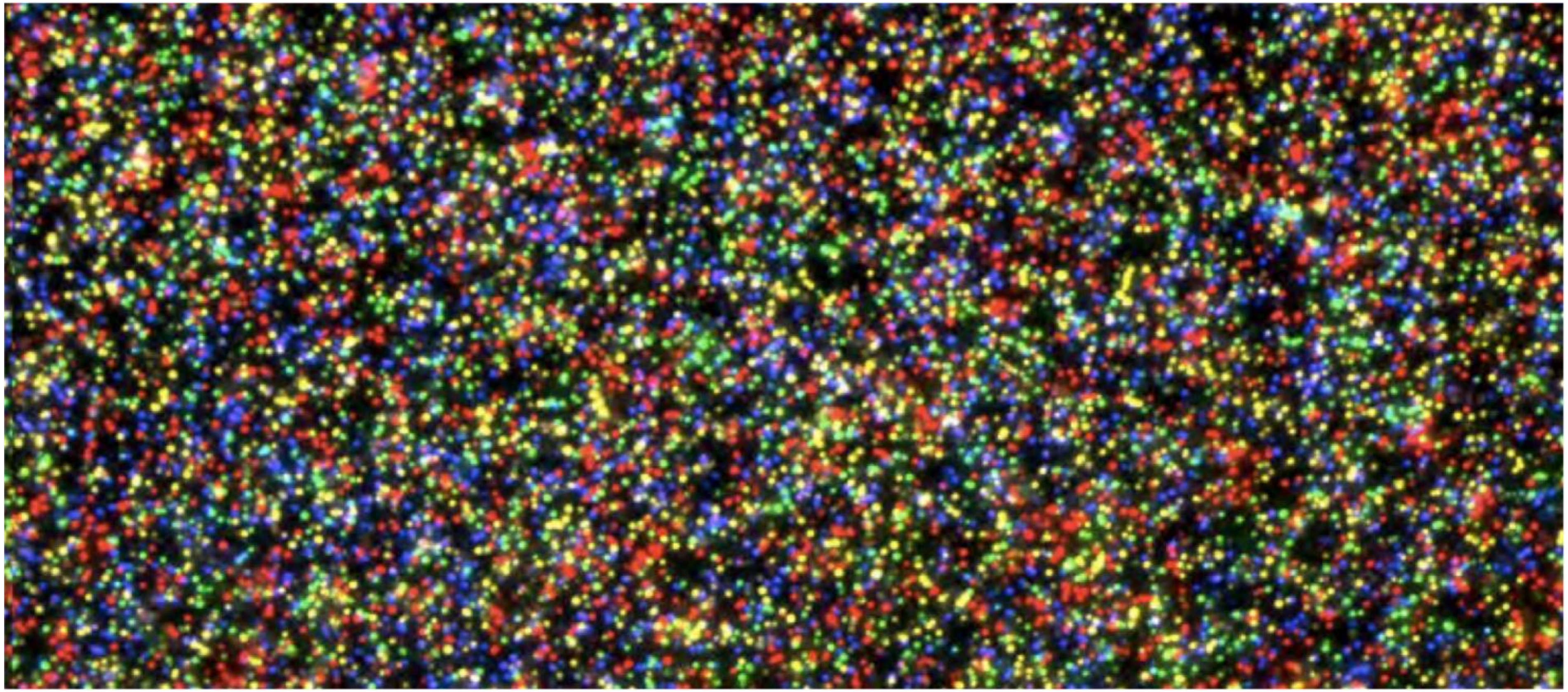
Base calling from raw data



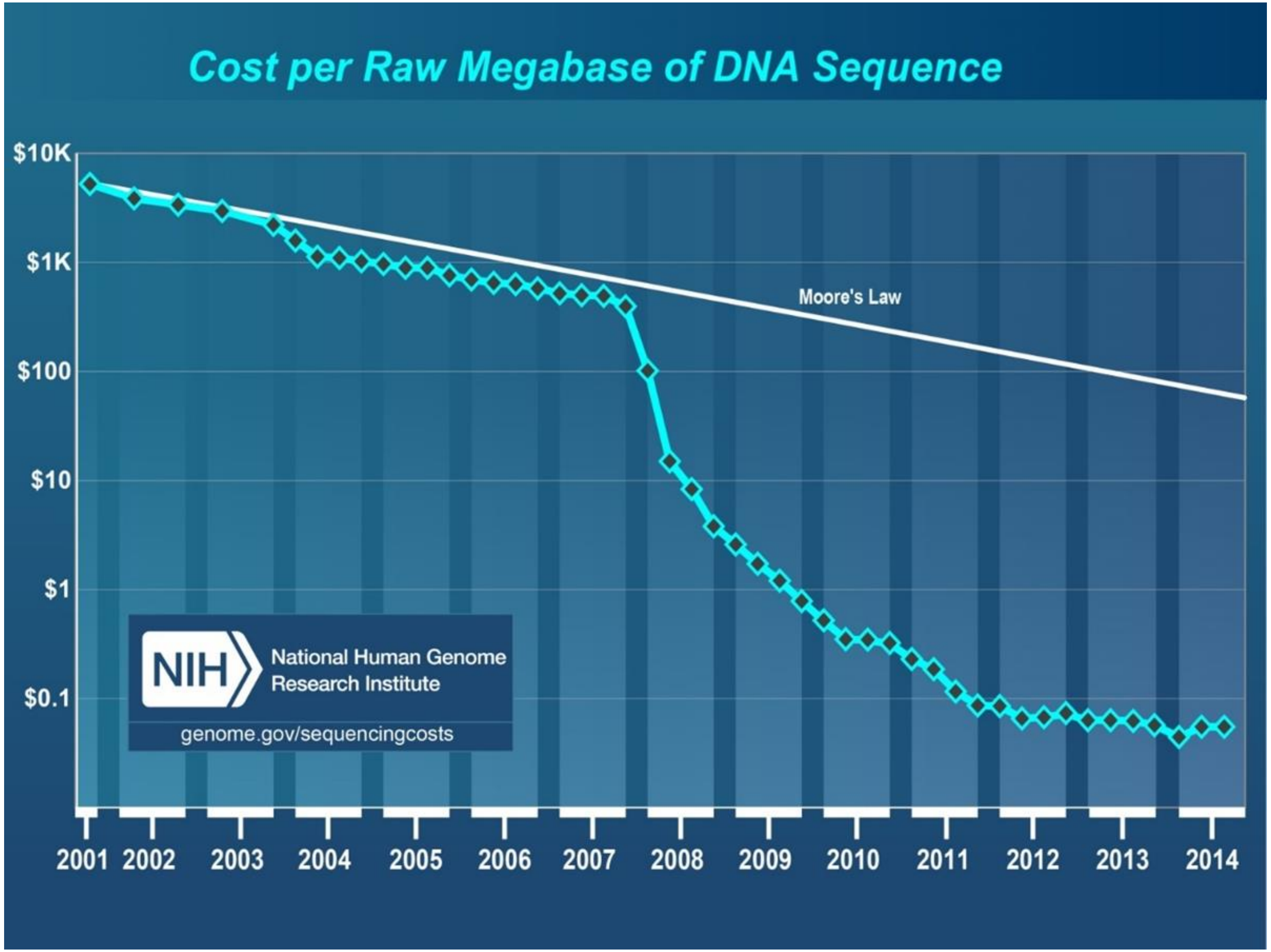
TGCTACGAT...

TTTTTTTGT...

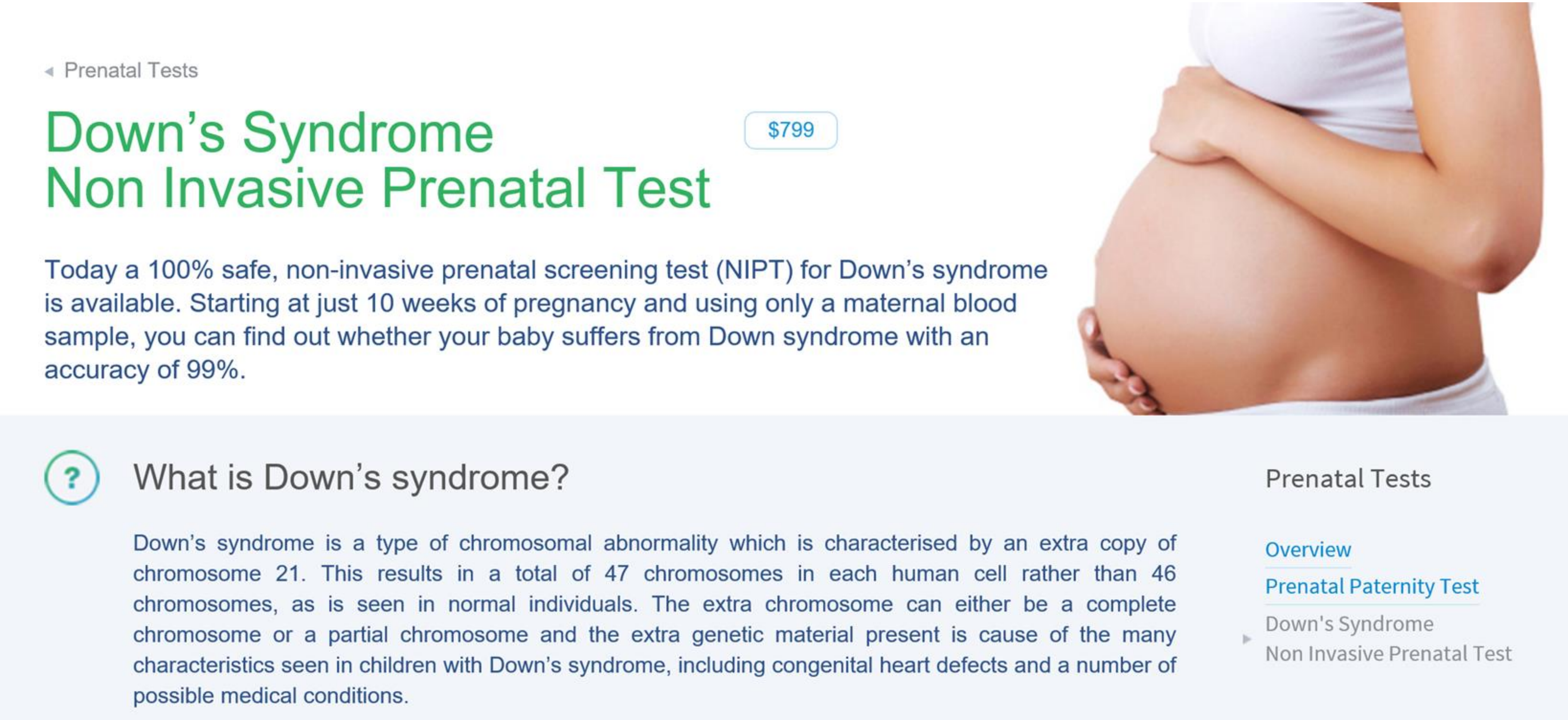
The identity of each base of a cluster is read off from sequential images



Bioinformatic analysis is essential !!


Cost per Raw Megabase of DNA Sequence
$10K
$1K
$100
$10
$1
$0.1
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014

Clinical Applications of Next Generation Sequencing Technologies



◂ Prenatal Tests

Down’s Syndrome Non Invasive Prenatal Test

$799

Today a 100% safe, non-invasive prenatal screening test (NIPT) for Down’s syndrome is available. Starting at just 10 weeks of pregnancy and using only a maternal blood sample, you can find out whether your baby suffers from Down syndrome with an accuracy of 99%.

? What is Down’s syndrome?

Down’s syndrome is a type of chromosomal abnormality which is characterised by an extra copy of chromosome 21. This results in a total of 47 chromosomes in each human cell rather than 46 chromosomes, as is seen in normal individuals. The extra chromosome can either be a complete chromosome or a partial chromosome and the extra genetic material present is cause of the many characteristics seen in children with Down’s syndrome, including congenital heart defects and a number of possible medical conditions.

Prenatal Tests

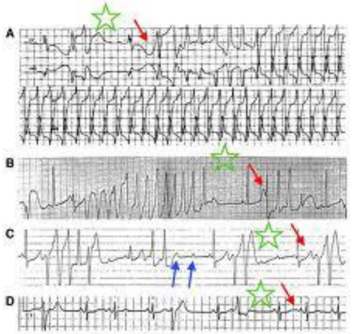
Overview
Prenatal Paternity Test
▸ Down's Syndrome Non Invasive Prenatal Test

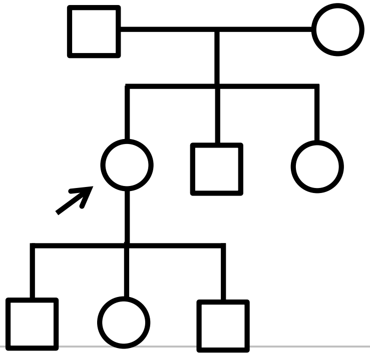
Potential Disease Gene Panels for Next Generation Sequencing

- Hypertrophic cardiomyopathy
- Dilated cardiomyopathy
- Hereditary arrhythmias (channelopathies)
- Retinitis pigmentosa
- Albinism
- Mental retardation
- Hearing loss

Heart disease

- 47 year female with sudden cardiac arrest
- Resuscitated successfully
- EKG reveals “Long QT Syndrome”
  - High risk for sudden death
  - Dozens of genes implicated
- Application of NGS to detect mutation
- Thereby guiding patient’s treatment and prevention of death in family members



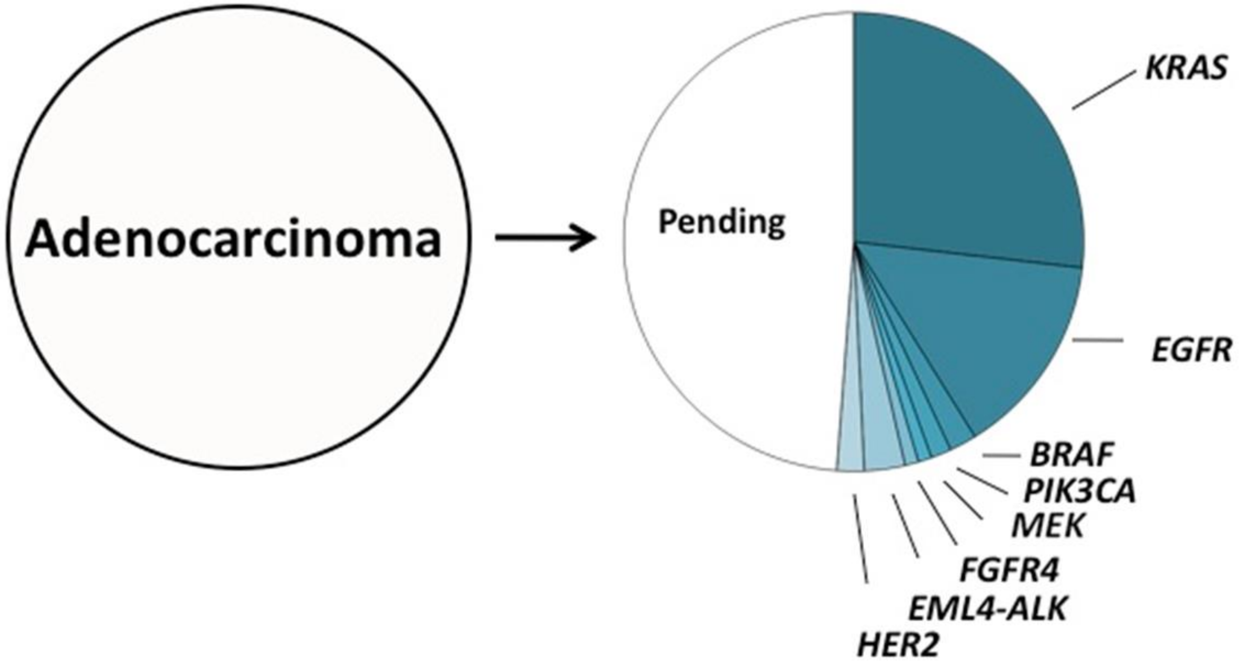


Cancer

Cancer is a heterogeneous disease that arises from accumulations of somatic mutations.

Next Generation Sequencing technologies share a fundamental process in which clonally amplified DNA templates, or single DNA molecules, are sequenced in a massively parallel fashion in a flow cell.

Molecular Profiling Can Explain The Heterogeneity of Lung Adenocarcinoma and Define Targets for Therapy

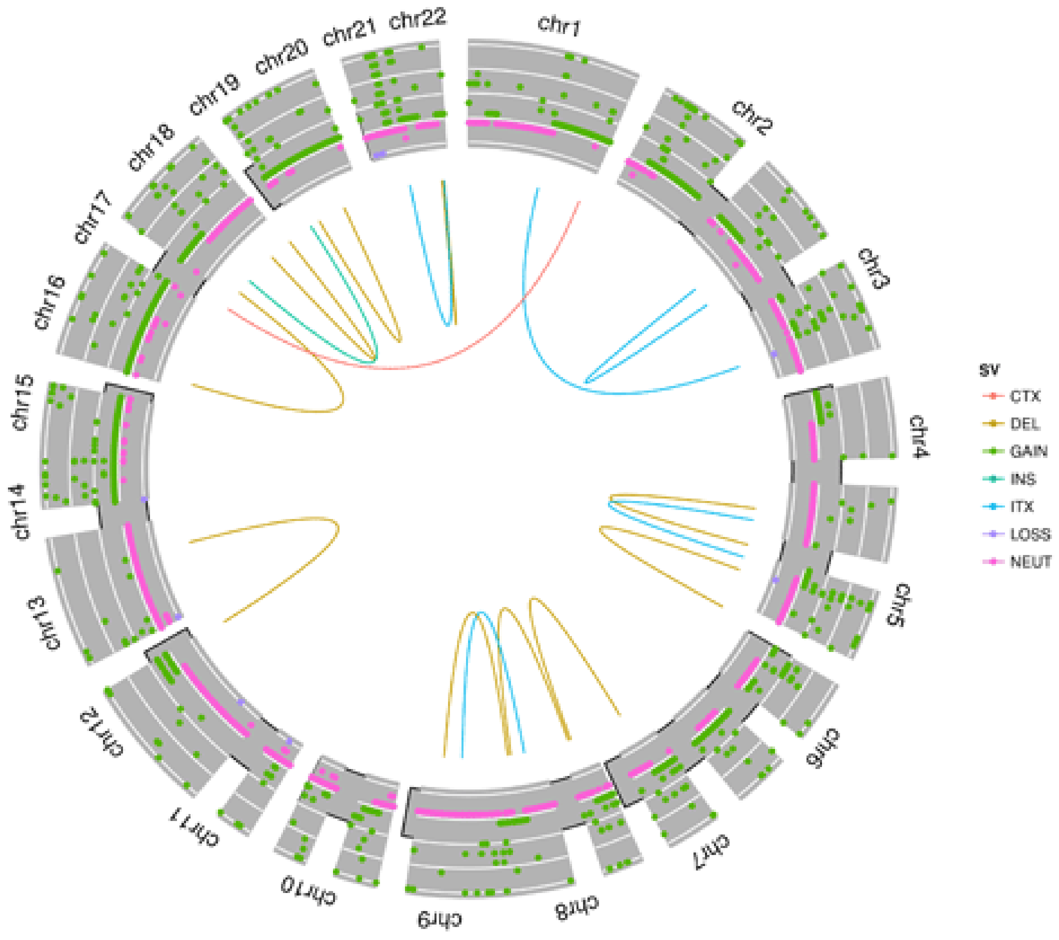


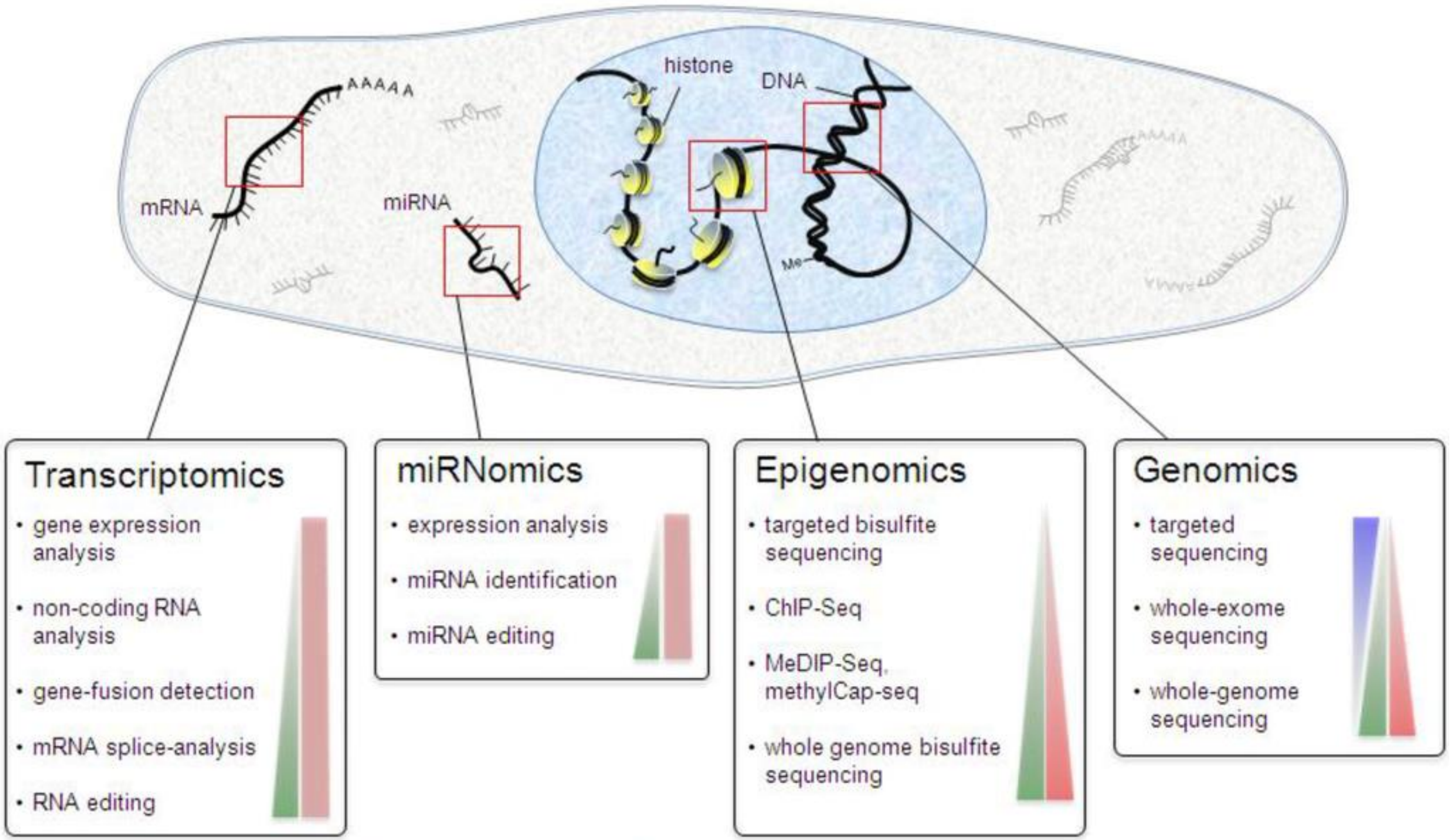
Targeted resequencing in molecular diagnostics

Targeted resequencing for somatic mutations serves as a useful and cost-effective platform to study a limited but relevant subset of putative cancer genes.

Fieuw A *et al.*, Cancer gene prioritization for targeted resequencing using FitSNP scores. PLoS One 2012;7:e31333.

Whole Genome Sequence of a Tumor





Transcriptomics

- gene expression analysis
- non-coding RNA analysis
- gene-fusion detection
- mRNA splice-analysis
- RNA editing

miRNomics

- expression analysis
- miRNA identification
- miRNA editing

Epigenomics

- targeted bisulfite sequencing
- ChIP-Seq
- MeDIP-Seq, methylCap-seq
- whole genome bisulfite sequencing

Genomics

- targeted sequencing
- whole-exome sequencing
- whole-genome sequencing

🟩 information content 🟥 experimental costs 🟦 current clinical utility

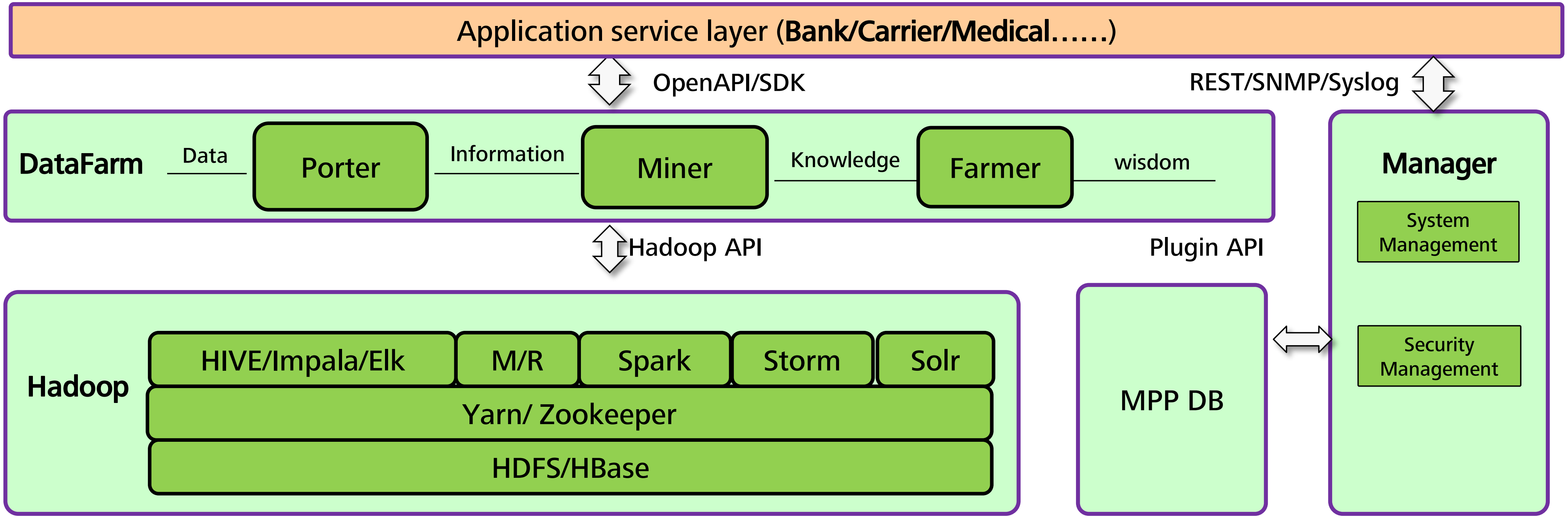
Part II

- **Huawei Big Data Introduction**

- **Bio Information/Medical Big Data Practices**

  1. **Gene sequencing analysis speedup – parallelization**
  2. **Co-Innovation with First Affiliated Hospital of Zhengzhou University**

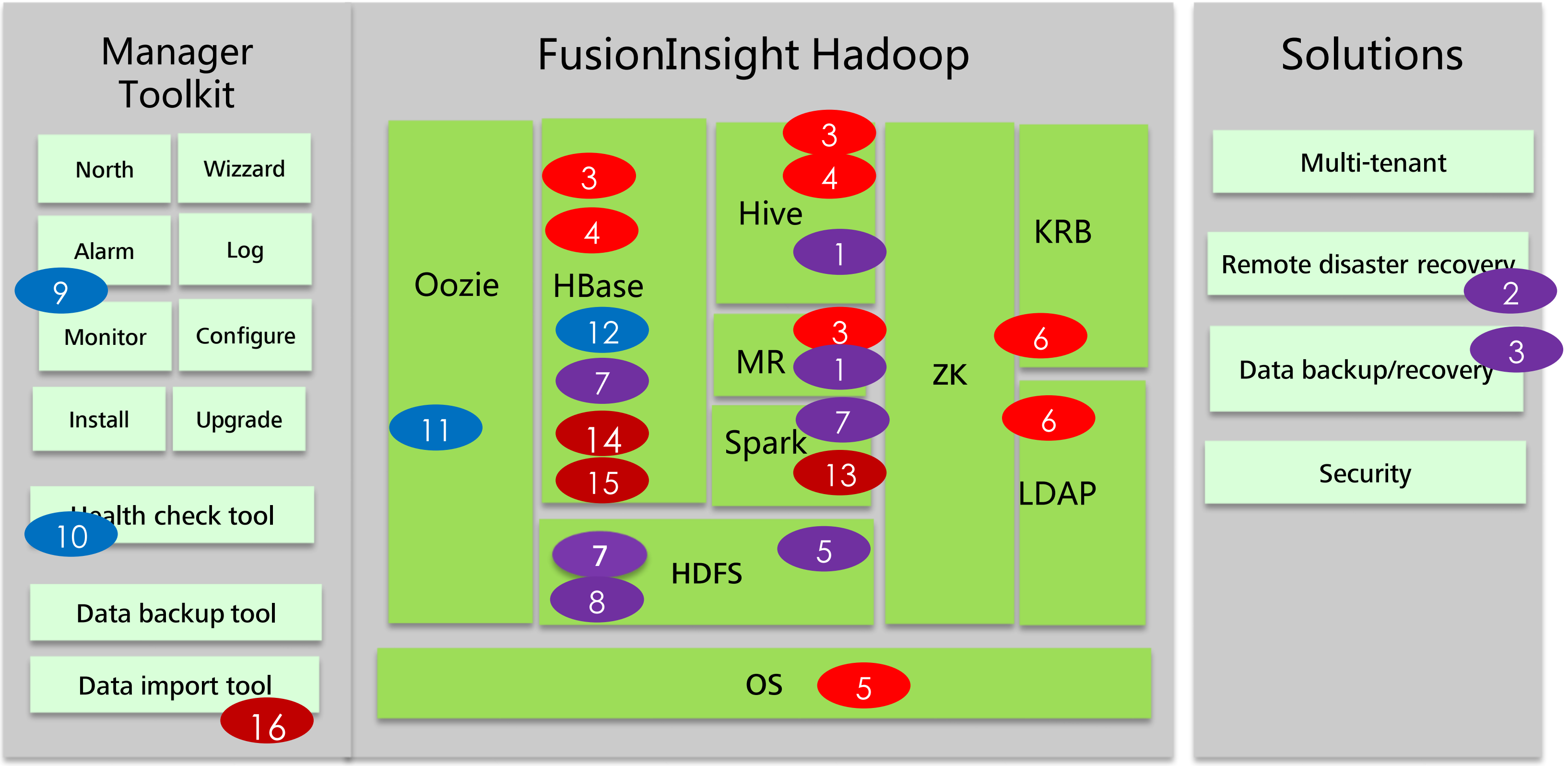
Huawei Big Data Product: FusionInsight Architecture



- **Hadoop layer:** enhancements based on open-source software and self R&D.
- **DataFarm layer**: end-to-end data insight with data integration service + data mining service + data service framework
- **Manager:** distributed system management framework with system management (OM/NTP/Disaster recovery), data security and governance.

FusionInsight Hadoop- Enterprise Level Secure, Reliable, Smart, Simple

Competitiveness	Key Optimization	No
Reliability	All Service HA	1
	Remote Disaster Recovery	2
	Data backup & recovery	3
	Service overload control	4
Security	OS Security Reinforcement	5
	Account management	6
	Right management based on accounts and roles	7
	Data encryption to protect	8
Simple	GUI interface, Installation wizard & upgrade tool	9
	Heath check and log collect tools	10
	Work flow	11
	SDK for Application Dev	12
Performance	MR task schedule algorithm optimization	13
	CTBase	14
	Secondary Index	15
	Parallel data import tool	16



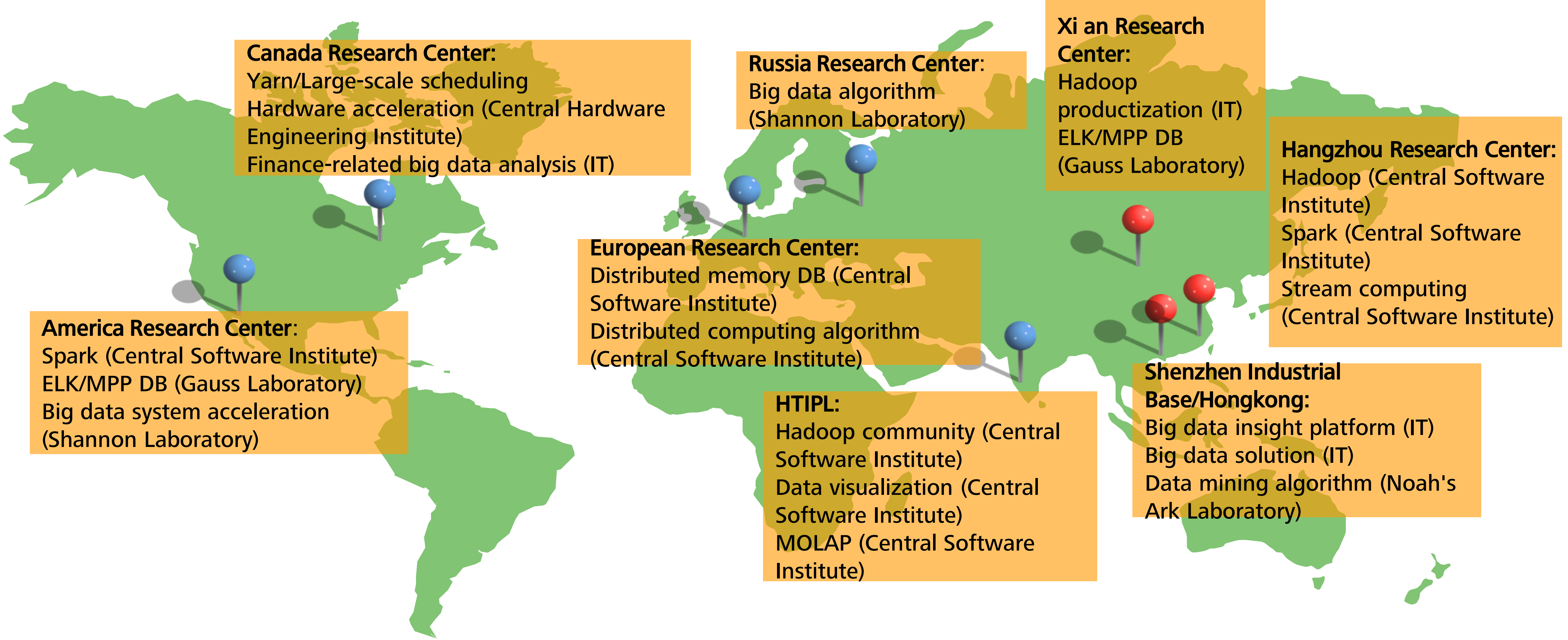
Based on open source for enterprise-class engineering optimization:

1. All components of physical healthy;

2. Application & management-oriented solutions and tools.

Principle

Global Layout and Full Coverage of Key Technologies in the Big Data Field



Canada Research Center:
Yarn/Large-scale scheduling
Hardware acceleration (Central Hardware Engineering Institute)
Finance-related big data analysis (IT)

Russia Research Center:
Big data algorithm (Shannon Laboratory)

Xi an Research Center:
Hadoop productization (IT)
ELK/MPP DB (Gauss Laboratory)

Hangzhou Research Center:
Hadoop (Central Software Institute)
Spark (Central Software Institute)
Stream computing (Central Software Institute)

European Research Center:
Distributed memory DB (Central Software Institute)
Distributed computing algorithm (Central Software Institute)

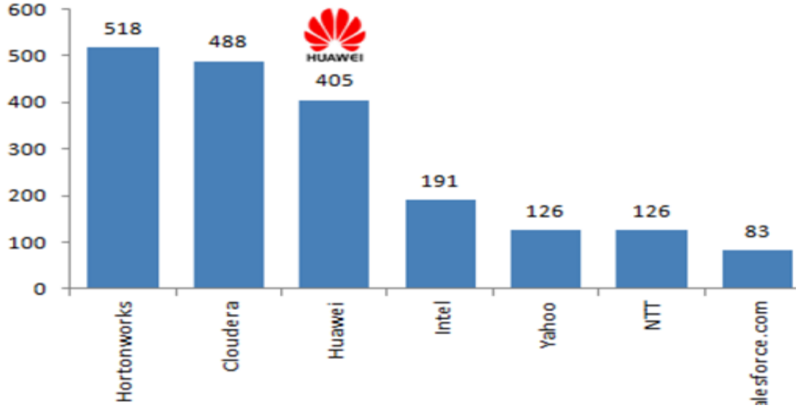
America Research Center:
Spark (Central Software Institute)
ELK/MPP DB (Gauss Laboratory)
Big data system acceleration (Shannon Laboratory)

HTIPL:
Hadoop community (Central Software Institute)
Data visualization (Central Software Institute)
MOLAP (Central Software Institute)

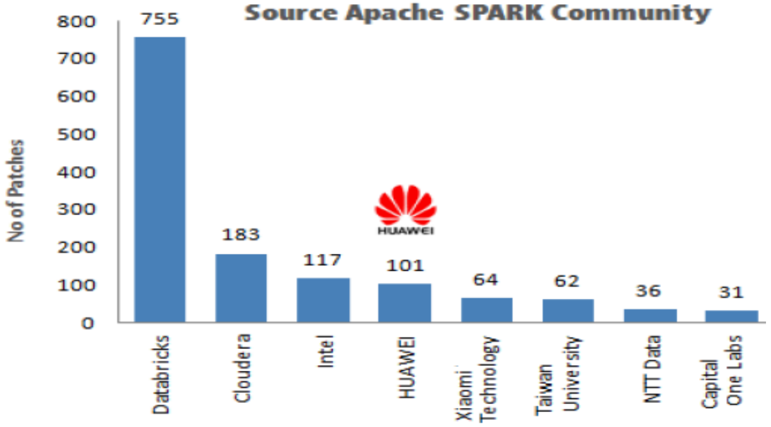
Shenzhen Industrial Base/Hongkong:
Big data insight platform (IT)
Big data solution (IT)
Data mining algorithm (Noah's Ark Laboratory)

Huawei is a Top Contributor to Apache Hadoop

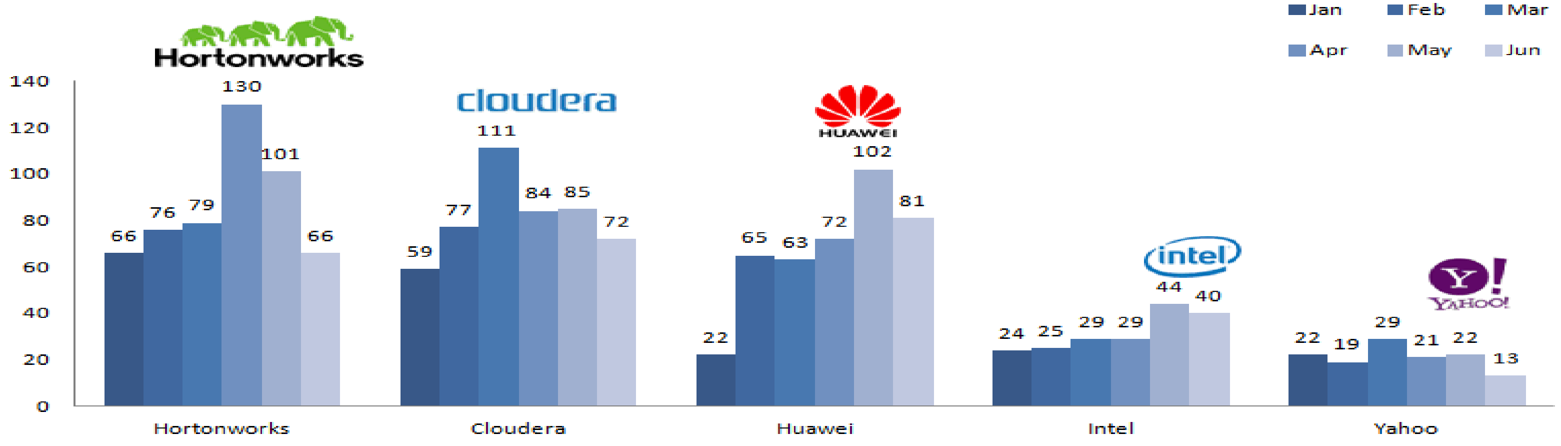
2015 - Huawei Contributions to Hadoop OS Community
(HDFS, YARN, MR, Hadoop Common, HBASE & Zookeeper)



2015 - HUAWEI Contributions to Open Source Apache SPARK Community



2015 - Top 5 Organization Contributions to Hadoop Open source Community Monthly Trend



Huawei master the nuclear core technique.

Huawei is in the first echelon of contributions.

Huawei's specialized contribution R&D team has 50+ (in Bangalore, India Big Data R&D group Team)

Main contribution:

- Secondary Index
- HBase Astro
- Spark Carbon

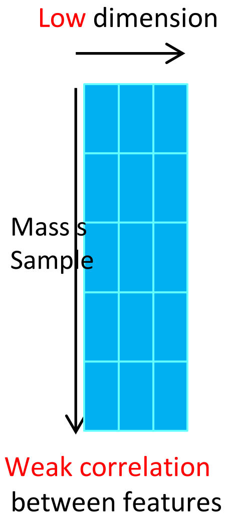
Main key feature:

- Ec code
- Spark SQL

Challenges of Gene Sequence Analysis: high dimension/super correlation/dense

Internet big data

Mass, Low dimension, Sparse, Weak correlation



※Low feature dimension

- ✓ Ten thousand level usually
- ✓ Usually by feature combination construct high-dimensional feature to improve precision of data mining
- ✓ Feature independently

※Data is huge, but Sparse

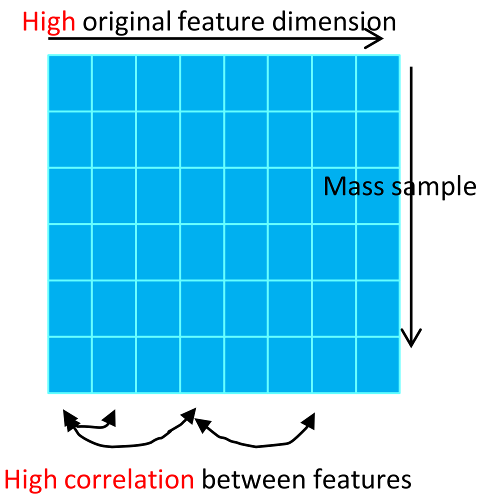
- ✓ e.g. Taobao， Sparseness:10e-08
- ✓ calculation and memory not too much

Computing requirements

ten thousands times

Gene big data

Mass, Super dimension, Dense, Strong correlation


High original feature dimension
Mass sample
High correlation between features

※ High feature dimension with Hundred million

- ✓ Characterized by the expression of the human protein
- ✓ 20,000 human genes
- ✓ Average 5 protein expression of each gene
- ✓ Each protein modification average 1000
- ✓ 20000*5*1000=》hundred million feature

※ Data is huge with Dense

- ✓ Gene: AGTC 4 kinds of base pairs，4^n
- ✓ Protein consists of 20 kinds of amino acids,20^n
- ✓ 1000 modification expressed in any position of 20^n

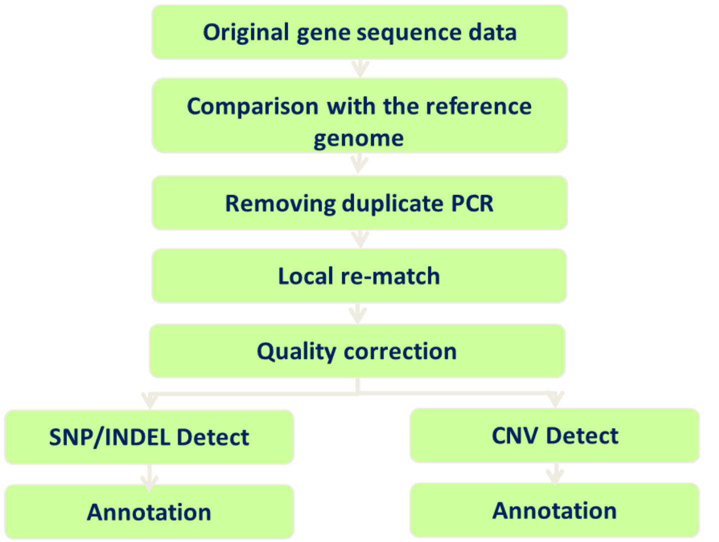
Normal Machine Learning

Medium IO, Weak compute, Weak overhead

Structure Machine Learning

Strong IO, Super computing, Super overhead

Gene sequencing analysis take much time – e.g. variation detection



Corresponding Software

BWA, Bowtie2, Soap, etc.

Picard, elPrep, SAMtools, etc.

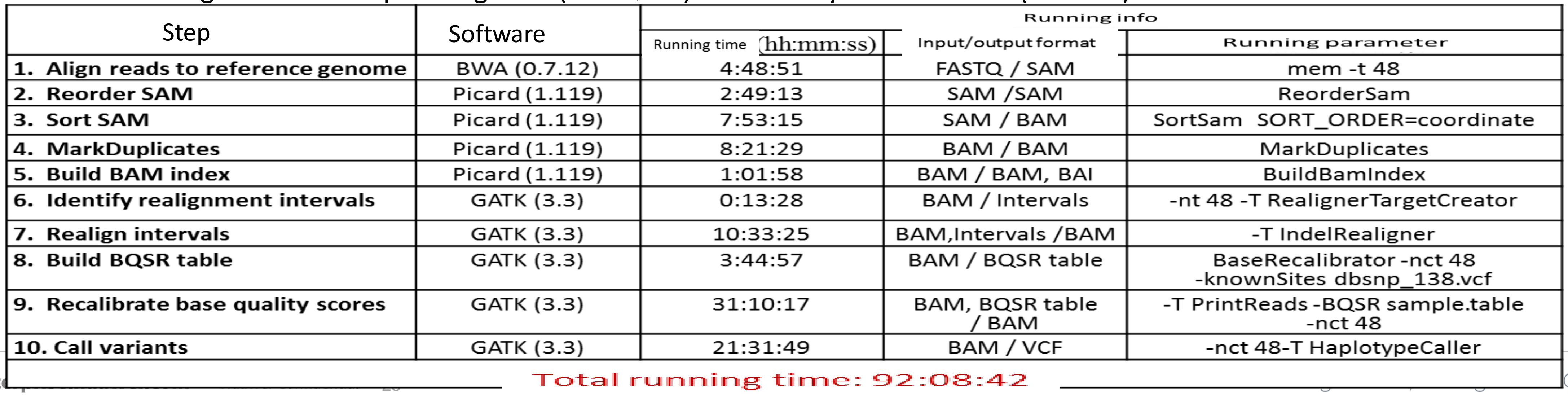
GATK, etc.

GATK, etc.

GATK，CNVnator，DNACopy，Varscan，SomaticIndel, etc.

- Support running on a single machine only, each operation using scripts concatenated
- Although support multithreading, resource utilization is low

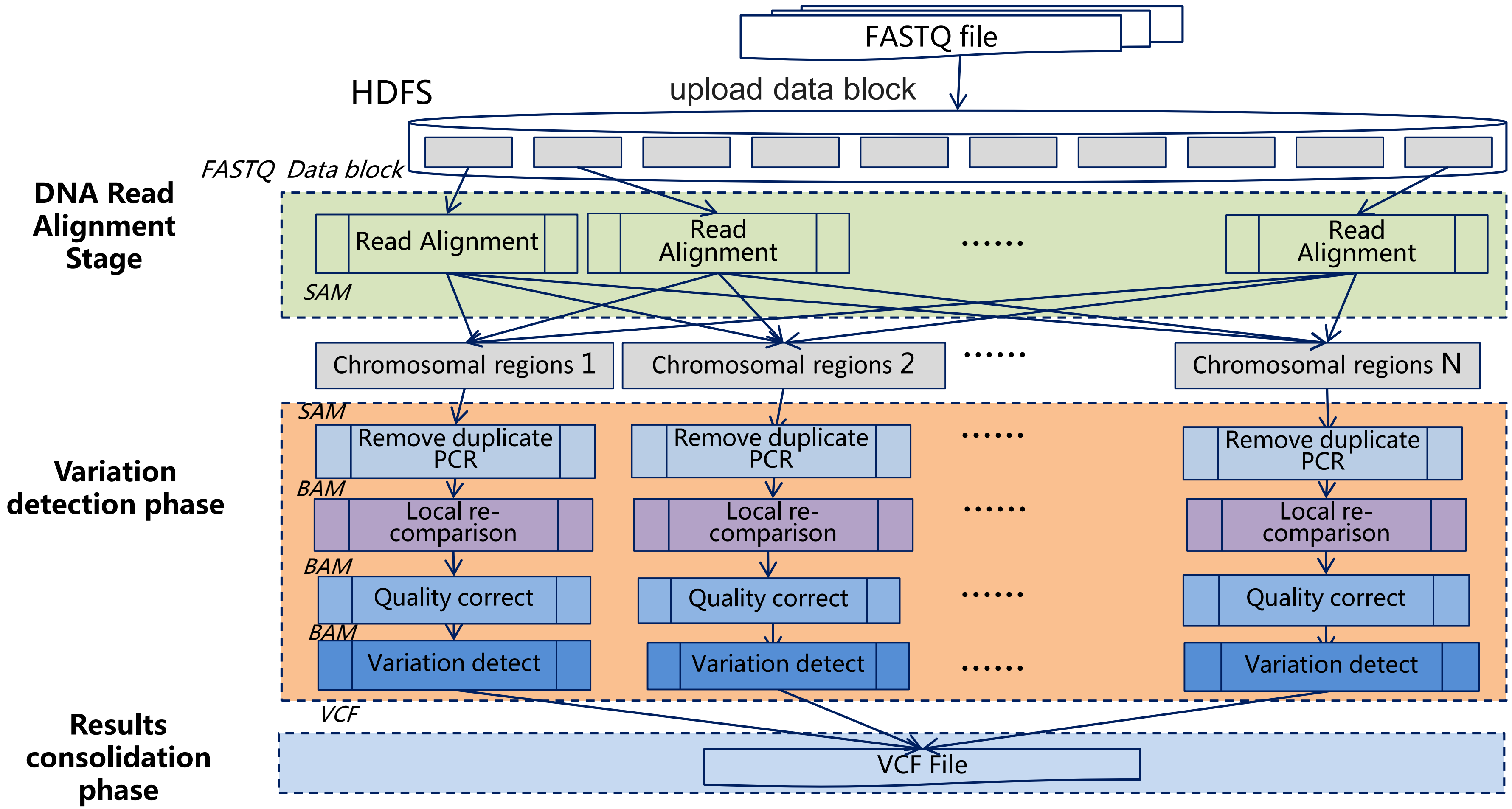
Origional DNA sequencing data (FastQ file)->Variability of test results(VCF file)



Step	Software	Running info		
		Running time (hh:mm:ss)	Input/output format	Running parameter
1. Align reads to reference genome	BWA (0.7.12)	4:48:51	FASTQ / SAM	mem -t 48
2. Reorder SAM	Picard (1.119)	2:49:13	SAM /SAM	ReorderSam
3. Sort SAM	Picard (1.119)	7:53:15	SAM / BAM	SortSam SORT_ORDER=coordinate
4. MarkDuplicates	Picard (1.119)	8:21:29	BAM / BAM	MarkDuplicates
5. Build BAM index	Picard (1.119)	1:01:58	BAM / BAM, BAI	BuildBamIndex
6. Identify realignment intervals	GATK (3.3)	0:13:28	BAM / Intervals	-nt 48 -T RealignerTargetCreator
7. Realign intervals	GATK (3.3)	10:33:25	BAM,Intervals /BAM	-T IndelRealigner
8. Build BQSR table	GATK (3.3)	3:44:57	BAM / BQSR table	BaseRecalibrator -nct 48 -knownSites dbsnp_138.vcf
9. Recalibrate base quality scores	GATK (3.3)	31:10:17	BAM, BQSR table / BAM	-T PrintReads -BQSR sample.table -nct 48
10. Call variants	GATK (3.3)	21:31:49	BAM / VCF	-nct 48-T HaplotypeCaller

Total running time: 92:08:42

Variation detection based on distributed parallel platforms - Acceleration

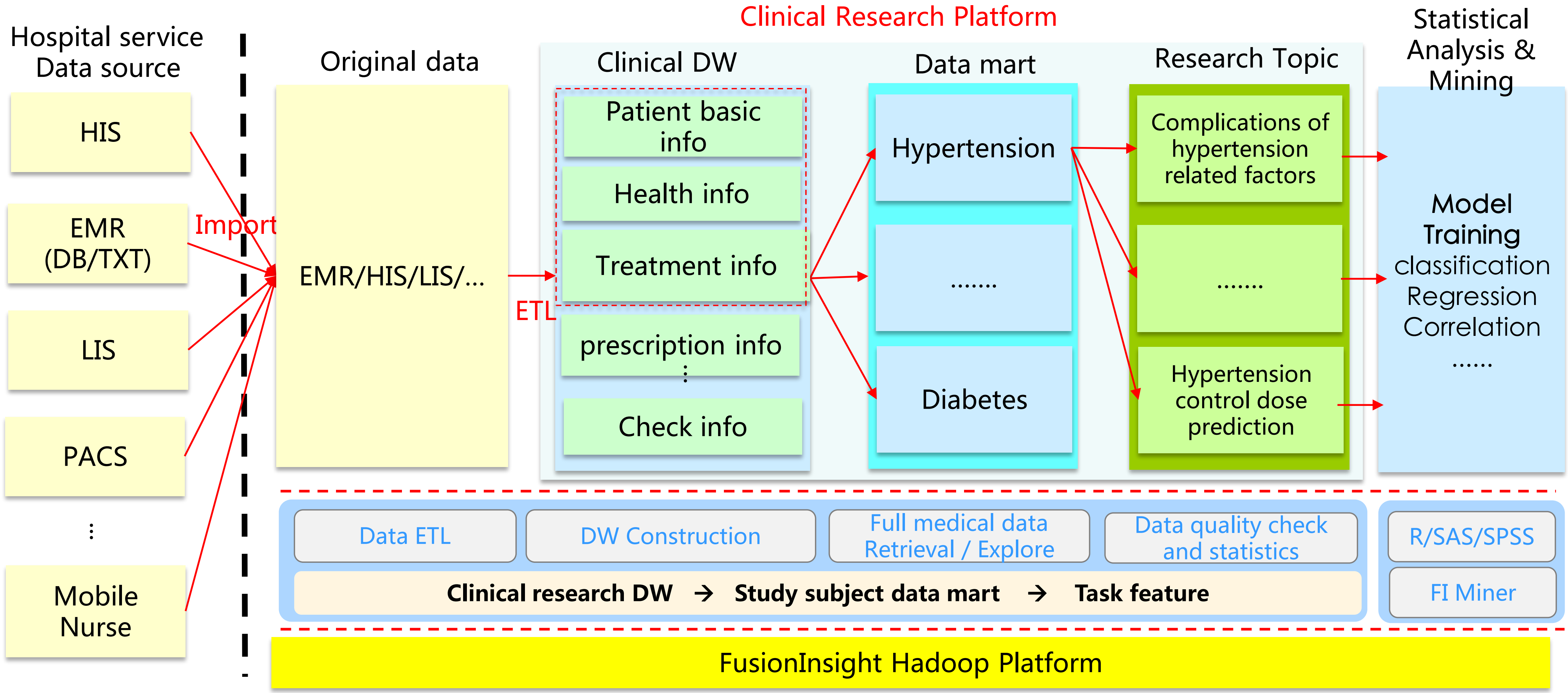


Key Issue

- ✓ Data partitioning
- ✓ Task Scheduling
- ✓ Resource allocation
- ✓ Load balance
- ✓ Global optimization

Acceleration Effect: time of whole gene analysis shortened from 92 to 1 hours ! All Exon-group analysis less than 10 minutes !

Huawei Helps to Build a Clinical Research Platform based on Hadoop



Clinical Research Platform

Hospital service Data source

- HIS
- EMR (DB/TXT)
- LIS
- PACS
- ⋮
- Mobile Nurse

Import

Original data

EMR/HIS/LIS/…

ETL

Clinical DW

- Patient basic info
- Health info
- Treatment info
- prescription info
- ⋮
- Check info

Data mart

- Hypertension
- …….
- Diabetes

Research Topic

- Complications of hypertension related factors
- …….
- Hypertension control dose prediction

Statistical Analysis & Mining

Model Training
classification
Regression
Correlation
……

Data ETL | DW Construction | Full medical data Retrieval / Explore | Data quality check and statistics

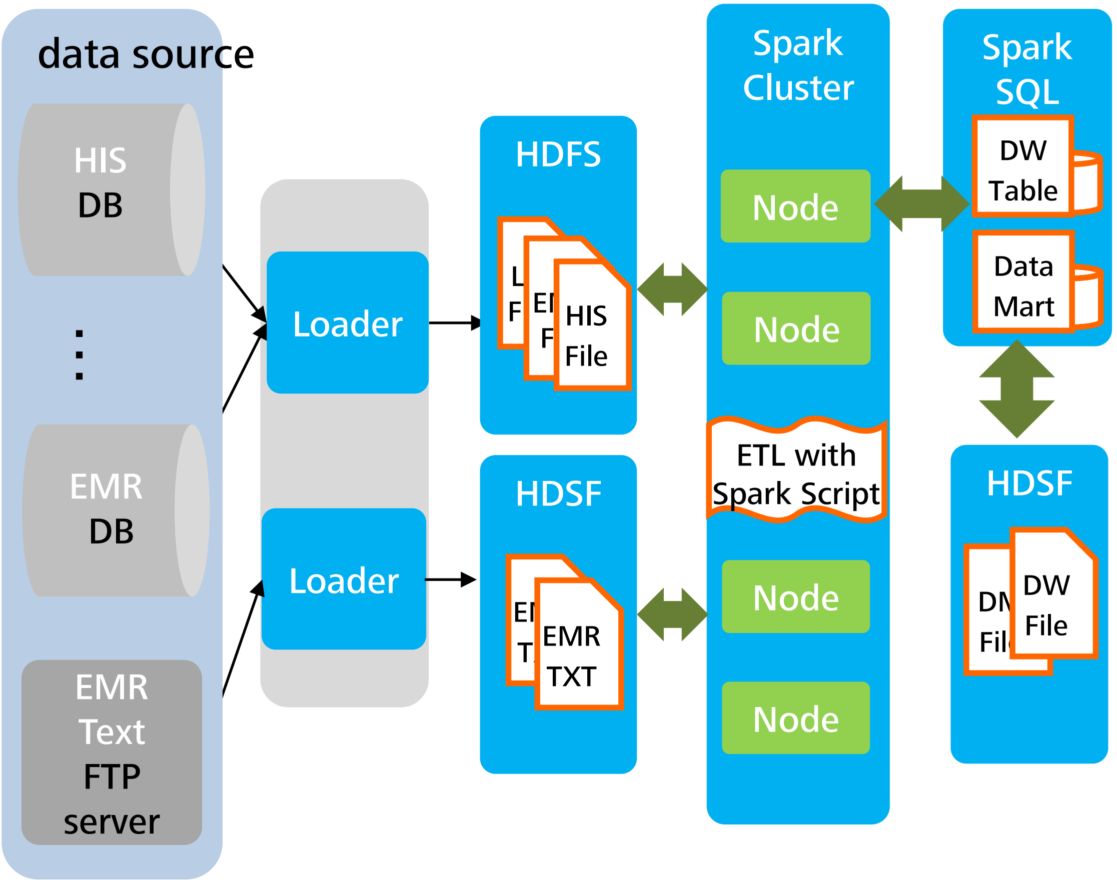
Clinical research DW → Study subject data mart → Task feature

R/SAS/SPSS

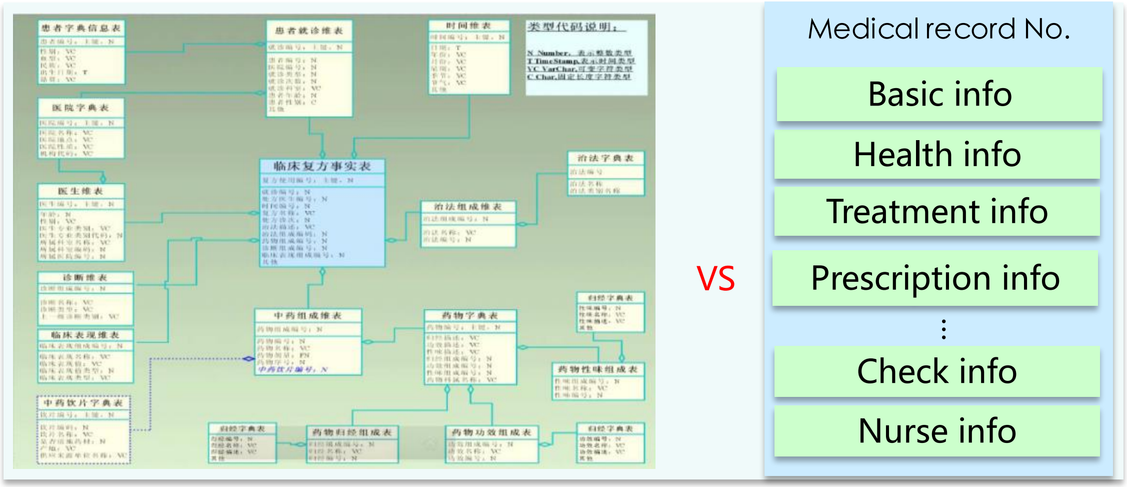
FI Miner

FusionInsight Hadoop Platform

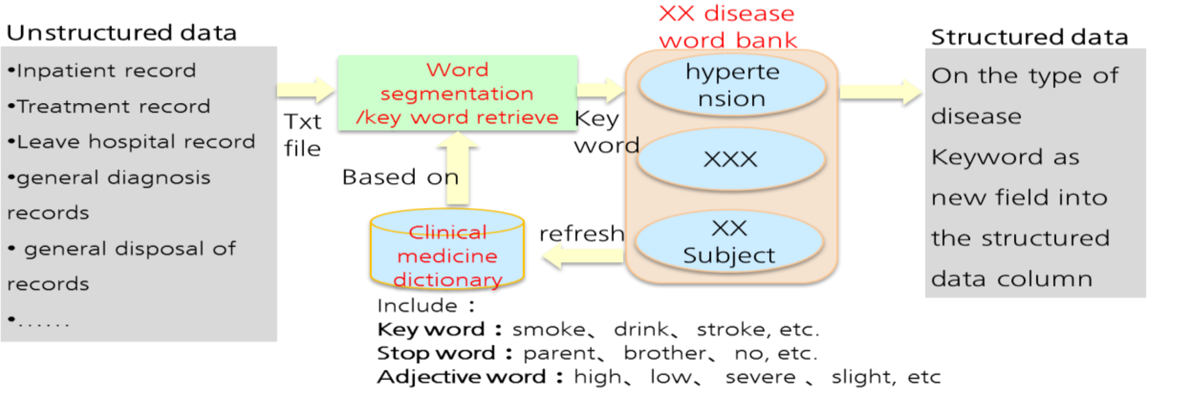
Clinical Research Platform Key technique



- 1.5 person-months
- billions of database record and 5 million electronic medical document
- almost 10 hypertension related study subject- analyze & model



Traditional model (star or snowflake) vs Big Data model (big table) clinical data warehouse schema comparison

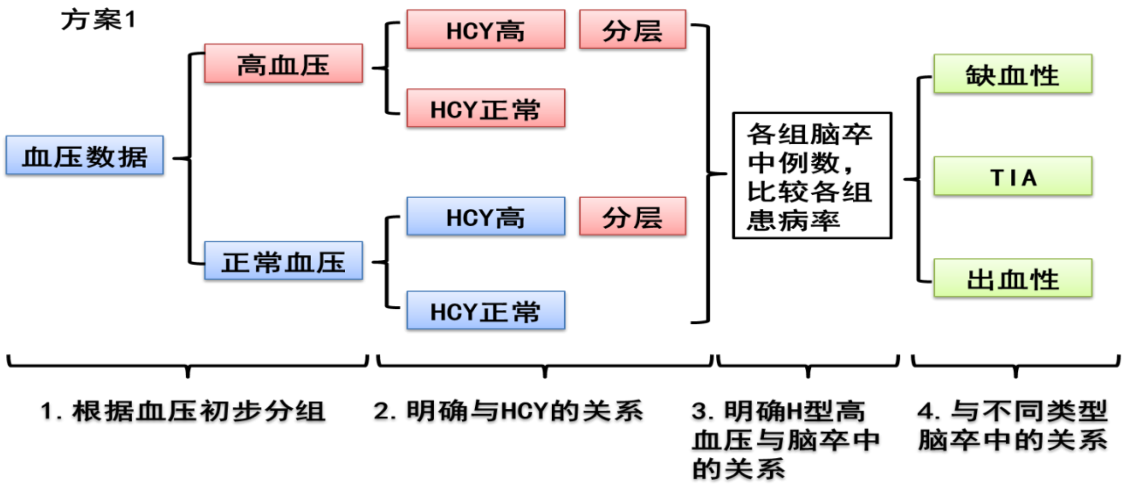

Unstructured data
•Inpatient record
•Treatment record
•Leave hospital record
•general diagnosis records
• general disposal of records
•……
Txt file
Word segmentation /key word retrieve
Based on
Clinical medicine dictionary
refresh
Key word
XX disease word bank
hypertension
XXX
XX Subject
Structured data
On the type of disease Keyword as new field into the structured data column
Include：
Key word：smoke、drink、stroke, etc.
Stop word：parent、brother、no, etc.
Adjective word：high、low、severe 、slight, etc

Unstructured EMR data structuring process

Clinical Research Subject Example

Task1 :

- Henan province people HCY hypertension incidence
- Relation between HCY hypertension and stroke
- Impact of intervention on stroke



方案1

血压数据

高血压

正常血压

HCY高

分层

HCY正常

HCY高

分层

HCY正常

各组脑卒中例数，比较各组患病率

缺血性

TIA

出血性

1. 根据血压初步分组

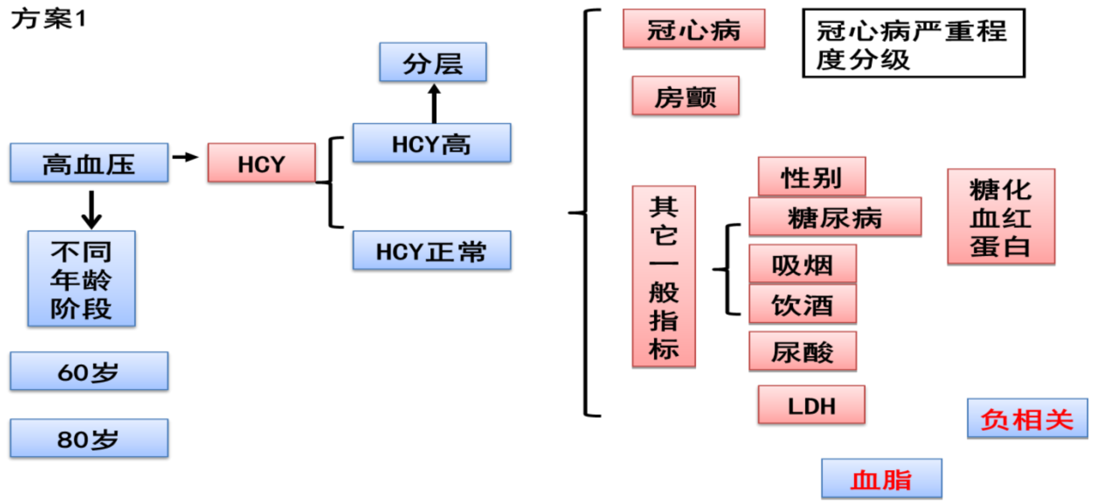
2. 明确与HCY的关系

3. 明确H型高血压与脑卒中的关系

4. 与不同类型脑卒中的关系

Taks2 :

- Henan province elderly hypertension patients observed with HCY level
- Relationship with coronary atherosclerosis



方案1

高血压

HCY

分层

HCY高

HCY正常

不同年龄阶段

60岁

80岁

冠心病

冠心病严重程度分级

房颤

其它一般指标

性别

糖尿病

吸烟

饮酒

尿酸

LDH

糖化血红蛋白

负相关

血脂

Other target organ damage analysis with HCY Hypertension:

- Coronary artery disease;
- Cardiac hypertrophy;
- Chronic kidney disease;
- Diabetes;
- Metabolic syndrome;
- Aortic dissection

Other impact factor:

- Age/Gender
- BMI Index
- Smoking
- Pulse pressure
- High/low density lipoprotein
- Carbamide/creatinine…….

THANK YOU

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.