High Throughput Sequencing (HTS)

Farhad Hormozdiari

UCLA

April 25, 2014

Genome Sequencing

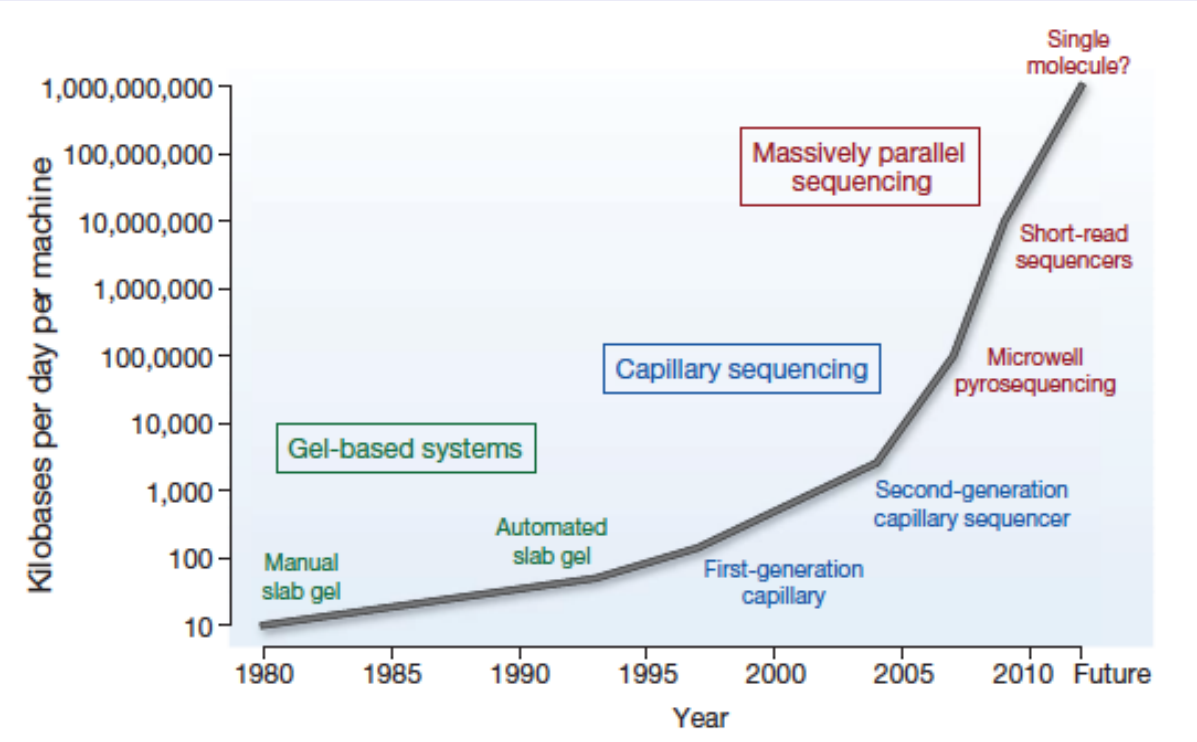
DNA

Strings of {A,C,T,G} *

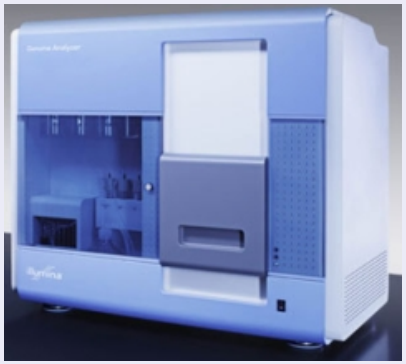
How?

- Sanger sequencing
- High throughput sequencing (HTS)
  - Illumina
  - 454
  - ABSolid
- Next generation sequencing (NGS)
  - PacBio

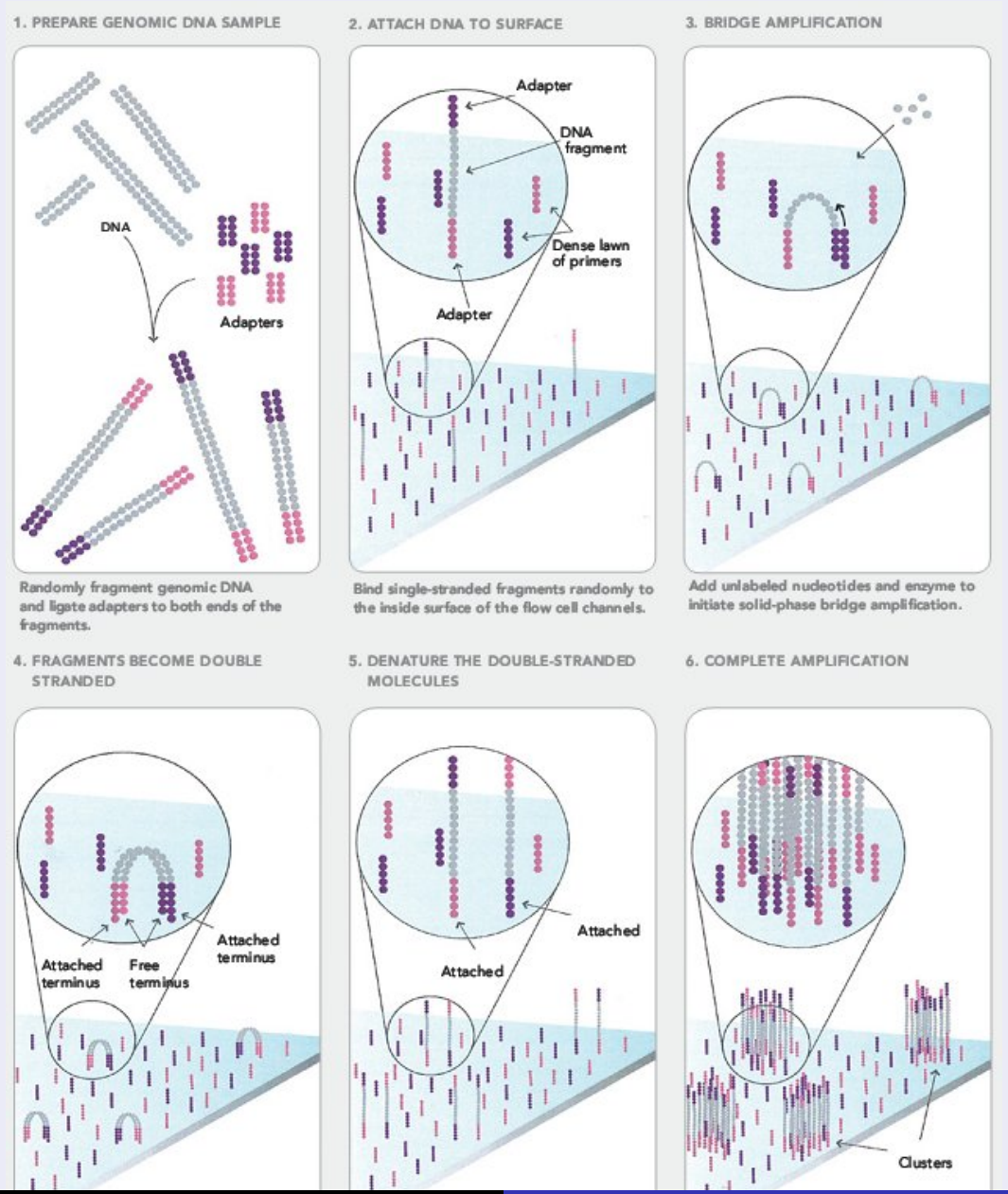
Illumina sequence generated in 30 years (Stratton et.al. 2009 *Nature Reviews*)



Illumina



ABolid


1. PREPARE GENOMIC DNA SAMPLE
DNA
Adapters
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
2. ATTACH DNA TO SURFACE
Adapter
DNA fragment
Dense lawn of primers
Adapter
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
3. BRIDGE AMPLIFICATION
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
4. FRAGMENTS BECOME DOUBLE STRANDED
Attached terminus
Free terminus
Attached terminus
5. DENATURE THE DOUBLE-STRANDED MOLECULES
Attached
Attached
6. COMPLETE AMPLIFICATION
Clusters

Time/Money vs. Accuracy/Length

Technology	Read length (bp)	Run time (days)	Gb per run	cost
Roche/454's	330	0.35	0.45	500K$
Illumina	36-100	4-9	18-35	540K$
SOLID3	25-50	7-14	30-50	595K$

Table: Metzker , Comparison of next-generation sequencing platforms, Nature Reviews 2010

How reads are generated?

>chr1
AACTGTGTCGTCGTGCGTACTCTCTACTACTACTACATCATCATA
AATCTTCTCTCTTTTCTCTTATATTATATAAAAAACCCCCCCCTT
ACGGTGTGTACATGCATAGACACTACGACTACAGTACGATGATAG
>chr1
AACTGTGTCGTCGTGCGTACTCTCTACTACTACTACATCATCATA
AATCTTCTCTCTTTTCTCTTATATTATATAAAAAACCCCCCCTT
ACGGTGTGTACATGCATAGACACTACGACTACAGTACGATGATAG

Read Generated1: TTTTCTCTTATATTATATAAAAAACC

Read Generated2: TGCATAGACACTACGACTACAGTACG

Whats the problem?

Detect the genetic variation between individuals?

Why?

- Drug design, different group of people *may* have different reaction to same drug.
- Disease study, why some individuals are more resistant to some diseases.

Solution

Detect the differences between the two sequence is a good map toward detecting the variation among individuals.

Heaven

Detect the genetic variation between individuals?

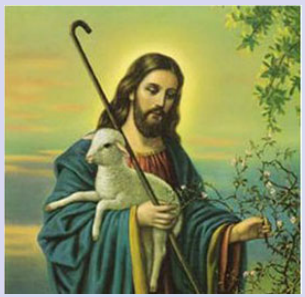

ACTGGGTATAAACTG


ATTGCGTATATACTG

Ideal case

We have?


ACTGGGTATAAACTG


ATT, *TGC*, *CGT*, *GTT*, *TAT*, *TAT*, *ATA*, *ACT*, *CTG*

Resequencing

Find where each read can align.

Use the alignment to detect the variation.

ACTGGGTTTAAACTG

ATT ✓

ATT ×

ATT ×

ATT ×

ATT ×

Assembly

$ATT, TGC, CGT, GTT, TAT, TAT, ATA, ACT, CTG$

Similarity Measure

Distance between two strings

Number of operations required to transfer one string to another string.

Hamming distance

Count number mistmaches between two strings.
Example:
string1: ACTCTTCCGT
string2: ACGCTCCGTA

Edit distance

Count number of indels or mismatches needed to transfer one string to another string.

Mapping

Hash-based

- Build hash table for the genome
  - Consider all possible 4^L, L length of hash key.
  - Record each occurrence in the genome.

Sequence	Positions
AAAAAAAAAA	32453, 64543, 76335
AAAAAAAAAC	64534, 84323, 96536
AAAAAAAAAG	12352, 32534, 56346
AAAAAAAAAT	23245, 54333, 75464
AAAAAAAACA	
AAAAAAAACC	43523, 67543
…	
CAAAAAAAAA	32345, 65442
CAAAAAAAAC	34653, 67323, 76354
…	
TCGACATGAG	54234, 67344, 75423
TCGACATGAT	11213, 22323
…	
TTTTTTTTTG	64252
TTTTTTTTTT	64246, 77355, 78453

BWA-based Trasformation

- Build the Burrows Wheeler Transform as mention in class is build for the genome
- searching is done by using the build index
- the index size is equal to genome size

Existing mappers

mr/mrsFAST	Bowtie	Maq	BWA
SHRiMP	SOAP	PerM	SOCS
ZOOM	BFAST	SSASHA	Novoalign
MOSAIK	RazerS	SLIDER	Stampy

mr/mrFAST (http://mrfast.sourceforge.net/)

Making index

./mrFAST –index genome.fa

>chr1
AACTGTGTCGTCGTGCGTNNNNNNNNN
AATCTTCTCTCTTTTCTCTTATATTATATA
or
>chr1
AACTGTGTCGTCGTGCGTNNNNNNNNN
AATCTTCTCTCTTTTCTCTTATATTATATA
>chr2
AACTGTGTCGTCGTGCGTNNNNNNNNN
AATCTTCTCTCTTTTCTCTTATATTATATA

Single-end mapping

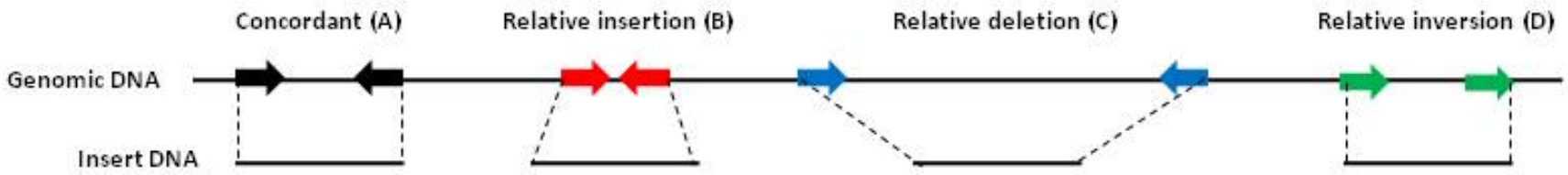
./mrFAST –search genome.fa –seq reads.fastq -o output -u unmap

It will generate two files:

- output: mapping positions for each reads.
- unmap: set of reads fail to map to genome.

Pair-end mapping

./mrFAST –search genome.fa –seq reads.fastq –pe –min 0 –max 400 –discordant-vh -o output -u unmap



Bowtie (http://sourceforge.net/projects/bowtie-bio/files/bowtie/0.12.7/)

Building index

bowtie-build genome.fa genome.index

Read mapping

bowtie genome.ebwt

- -a Report all valid alignments per read or pair.
- -v <int> Report alignments with at most <int> mismatches
- -q/-f The query input files fastq/fasta
- –best in terms of number of mismatches and quality

BWA (http://bio-bwa.sourceforge.net/)

Building index

bwa index genome.fa

Read mapping

bwa aln -n 6 genome.fa reads.fastq > out.sai

bwa samse -n 6 genome.fa out.sai reads.fastq > out.sam

- -o Maximum number of gap opens
- -t Number of threads (multi-threading mode)
- -n Maximum edit distance

Input

FASTQ

@ILLUMINA-96BC32-0001:1:1:27:942
TTAGACTTCACACACTTGCTAAGAGATTGCAATAAGAAACCTAATT
+ILLUMINA-96BC32-0001:1:1:27:942
ggegcgggggggggdggggegcggggdgggeadgddereceggegggggffdeegg

- read-name
- read-sequence
- +read-name
- read-quality

FASTA

>ILLUMINA-96BC32-0001:1:1:27:942
TTAGACTTCACACACTTGCTAAGAGATTGCAATAAGAAACCTAATT

- read-name
- read-sequence

Comparison Between mappers

Mapper	Platform	Distance	License
BFAST	Illumina/SOLiD	Edit	Downloadable
Bowtie	Illumina/SOLiD	Hamming	Downloadable
BWA	Illumina/SOLiD	Edit	Downloadable
MAQ	Illumina/SOLiD	Hamming	Downloadable
mrFAST	Illumina	Edit	Downloadable
mrsFAST	Illumina	Hamming	Downloadable
SHRiMP	Illumina/SOLiD	Edit	Downloadable
ZOOM	Illumina	Edit	Commercial

Comparison Between mappers

Hash-based

- BFAST
- PerM
- SHRiMP
- ZOOM
- mr/mrsFAST

BWT

- Bowtie
- BWA

Full sensitivity

If a read maps to specific locus it will find it.

- BFAST
- PerM
- SHRiMP
- ZOOM
- mr/mrsFAST

Before using a mapper know its limitations and strength.
For Example:
Not all mapper can map any length of reads.
Some mapper have upper bound on number of errors allowed in a read.

Mapper	read length
BFAST	NA
Bowtie	NA
BWA	NA
MAQ	50bp
mrFAST	NA
mrsFAST	NA
SHRiMP	NA
ZOOM	40bp

Parallization

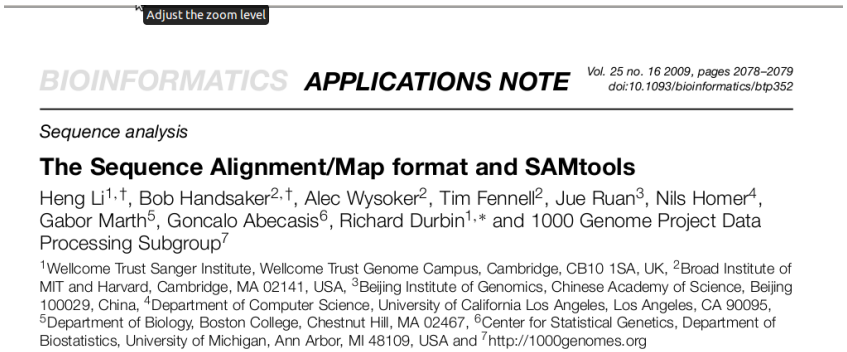
- Some mapper handle it by default
- If not just split the reads and give each splited chunk to one CPU.

SAM format

- read-name
- mapping-flag
- reference name
- mapping position
- mapping quality
- CIGAR string
- reference name of mate-read
- mapping position of mate-read
- number of errors
- MD field

SRR059723.6746268 0 chr10 3922082 255 36M * 0 0
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA IIIIIIIIIEIIEIIFHH
HGIFF¡@:¿@¿¡I;75C@ NM:i:2 MD:Z:3CC31

SAM tools (http://samtools.sourceforge.net/)



Adjust the zoom level

BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 16 2009, pages 2078–2079
doi:10.1093/bioinformatics/btp352

Sequence analysis

The Sequence Alignment/Map format and SAMtools

Heng Li[1,†], Bob Handsaker[2,†], Alec Wysoker[2], Tim Fennell[2], Jue Ruan[3], Nils Homer[4], Gabor Marth[5], Goncalo Abecasis[6], Richard Durbin[1,*] and 1000 Genome Project Data Processing Subgroup[7]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, [3]Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China, [4]Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, [5]Department of Biology, Boston College, Chestnut Hill, MA 02467, [6]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and [7]http://1000genomes.org

Usages

- Call SNPs/indels
- Manipulate the SAM files (i.e. convert it to BAM)

Convert SAM to BAM

- samtools faidx ref.fa
- samtools view -bt ref.fa.fai aln.sam > aln.bam
- samtools sort aln.bam aln-sorted
- samtools index aln-sorted.bam

IGV (http://www.broadinstitute.org/software/igv/)

Integrative genomics viewer

- Load genome and mapping file into viewer

How?

Load reference and BAM mapping file to IGV

IGV


a
Effect of MSA on alignments
NA12878, chr. 1:1,510,530–1,510,589
rs28782535
rs28783181
rs28788974
rs34877486
rs28788974
1000 Genomes Pilot 2 data, raw MAQ alignments
1000 Genomes Pilot 2 data, after MSA
b

TopHat (http://tophat.cbcb.umd.edu/)

Designed by Bowtie developers.

Why?

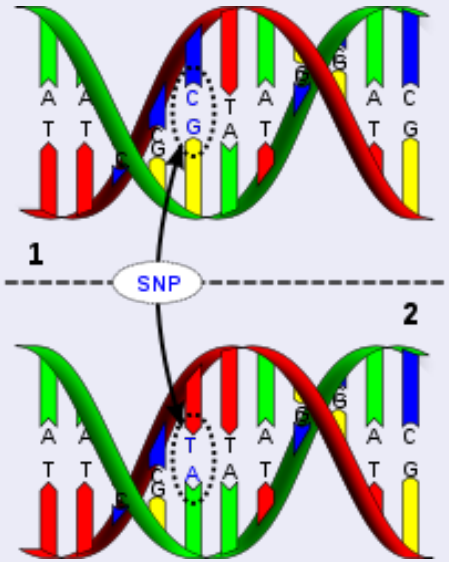
ref: AACACACACATTTATCTCTCTCCTCTCTCTCTTATATAATTA
read: AACACACACATATATAATTA
split:AACACACACA TATATAATTA

When?

- Breakpoint junctions (i.e. begining of CNVs)
- Junctions between exons in RNA-Seq

SNP call

Detecting SNP from short reads?



SNP caller

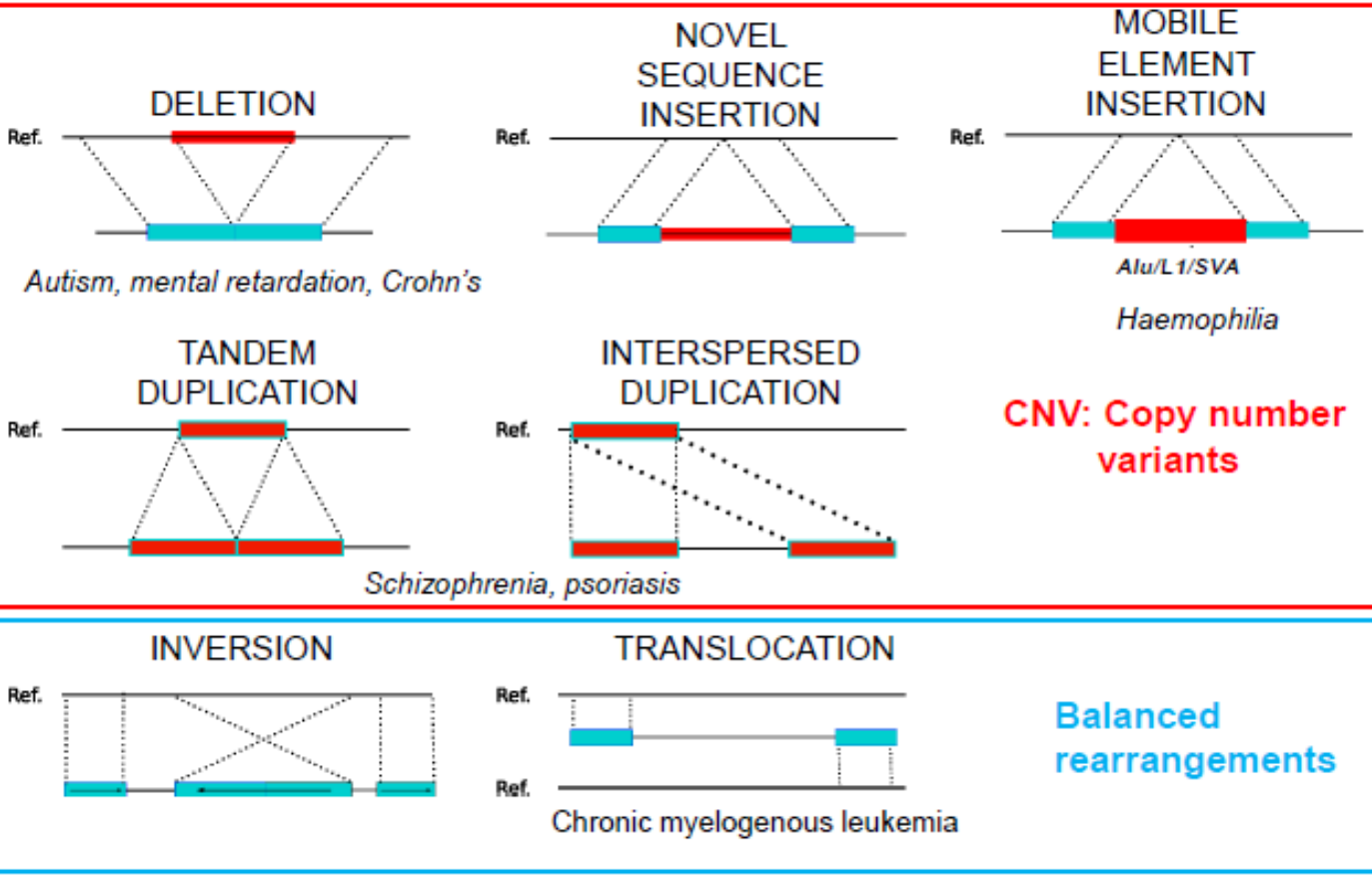
Algorithm	Platform	Strategy	SNP	Indel
MAQ	Illumina	Read pileup	YES	YES
GATK	Illumina/SOLiD	Likelihood optimization	YES	NO
SAMtools	Illumina/SOLiD	Read pileup	YES	YES
SOAPsnp	Illumina/454	Likelihood optimization	YES	YES
Dindel	Illumina	Expection Maximization	NO	YES
Pindel	Illumina	Split read mapping	NO	YES
VARiD	Illumina/SOLiD	Probabilistic model (HMM)	YES	NO

Strcutural Variation:

Variation between two individuals can be larger than one base (SNPs)

- Insertion
- Deletion
- Inversion
- Translocation
- Copy Number Variation

Structural Variation Classes



Structural Variation(SV) detection methods

- VariationHunter
- BreakDancer
- CREST
- MODiL
- Hydra

SV callers

Algorithm	Strategy
Variationhunter	Discordant reads and clustering
BreakDancer	Pair-end reads and Probabilistic model
CREST	Splitting the reads
MODiL	Probabilistic model (EM) and clustering
Hydra	Similar to VariationHunter

Copy Number Variation methods

- mrCaNaVar
- CNVnator
- CNVer
- ExonCNV

CNV callers

Algorithm	Strategy
mrCaNaVar	Read depth
CNVnator	Read depth
CNVer	Discordant reads and Read depth
ExonCNV	Read depth

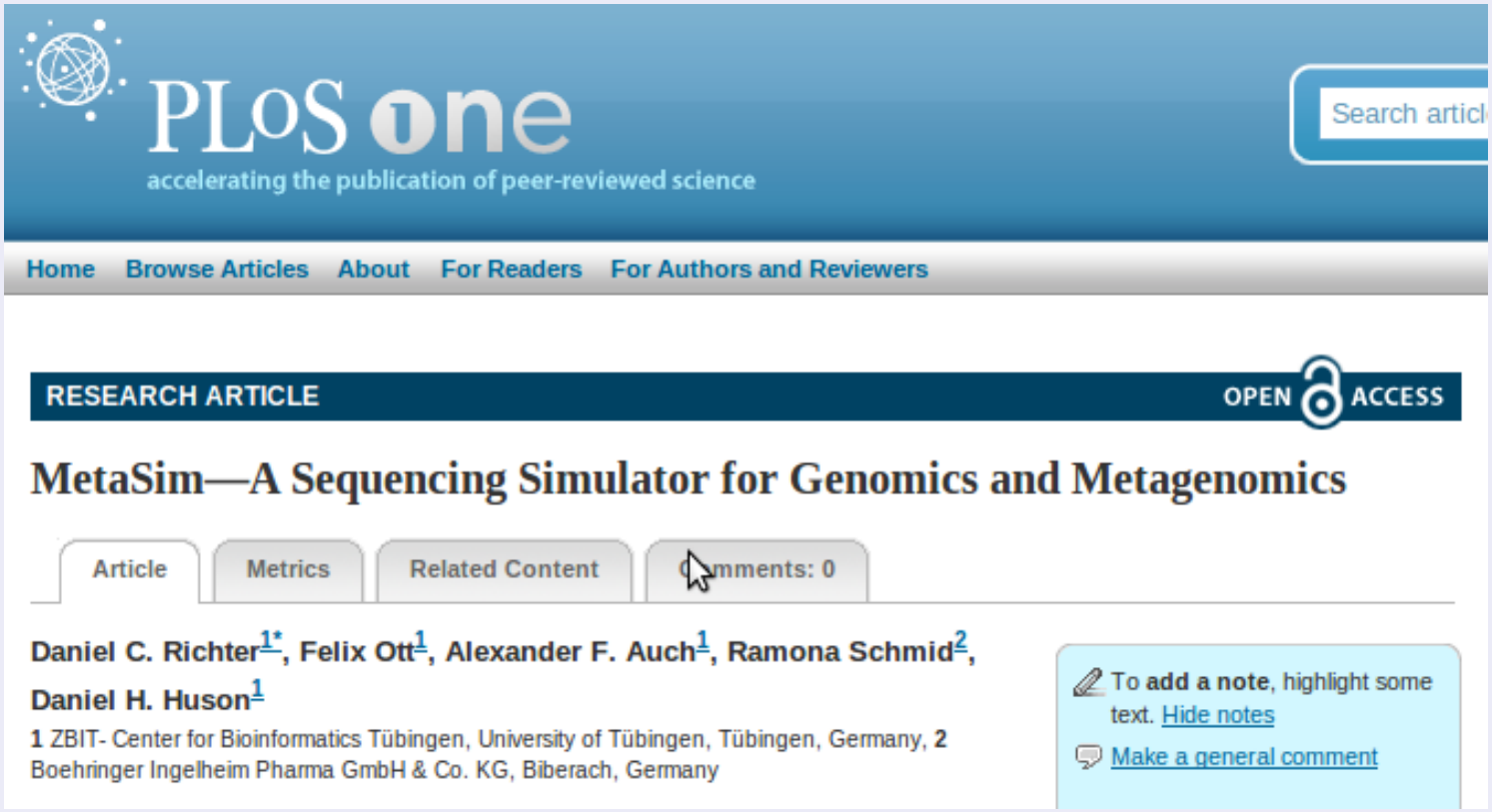
Read Simulating

How to generate sample reads?

Implement

Randomly select a position in the genome and return L bp following that positions.
Easy to implement.

- How to generate reads with Illumina like errors?
- How to generate pair-end reads with illumina like insert size?



PLoS one
accelerating the publication of peer-reviewed science

Search articl

Home Browse Articles About For Readers For Authors and Reviewers

RESEARCH ARTICLE OPEN ACCESS

MetaSim—A Sequencing Simulator for Genomics and Metagenomics

Article Metrics Related Content Comments: 0

Daniel C. Richter[1*], Felix Ott[1], Alexander F. Auch[1], Ramona Schmid[2], Daniel H. Huson[1]

1 ZBIT- Center for Bioinformatics Tübingen, University of Tübingen, Tübingen, Germany, 2 Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany

To **add a note**, highlight some text. Hide notes

Make a general comment

De novo Assembler

- EULER
- ABySS
- ALLPATHS-LG
- Celera
- IDBA
- MIRA
- TIGR
- SOAPdenovo
- Velvet

Galaxy (http://main.g2.bx.psu.edu/)

Galaxy

Framework which contains sets of tools for HTS.

- Easy to use
- Run small sets of data