

NGS Data Analysis

Roberto Preste

Useful info

Contacts:

roberto.preste@gmail.com

Slides:

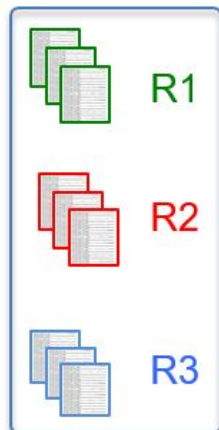
<http://bit.ly/ngs-data>

NGS data analysis

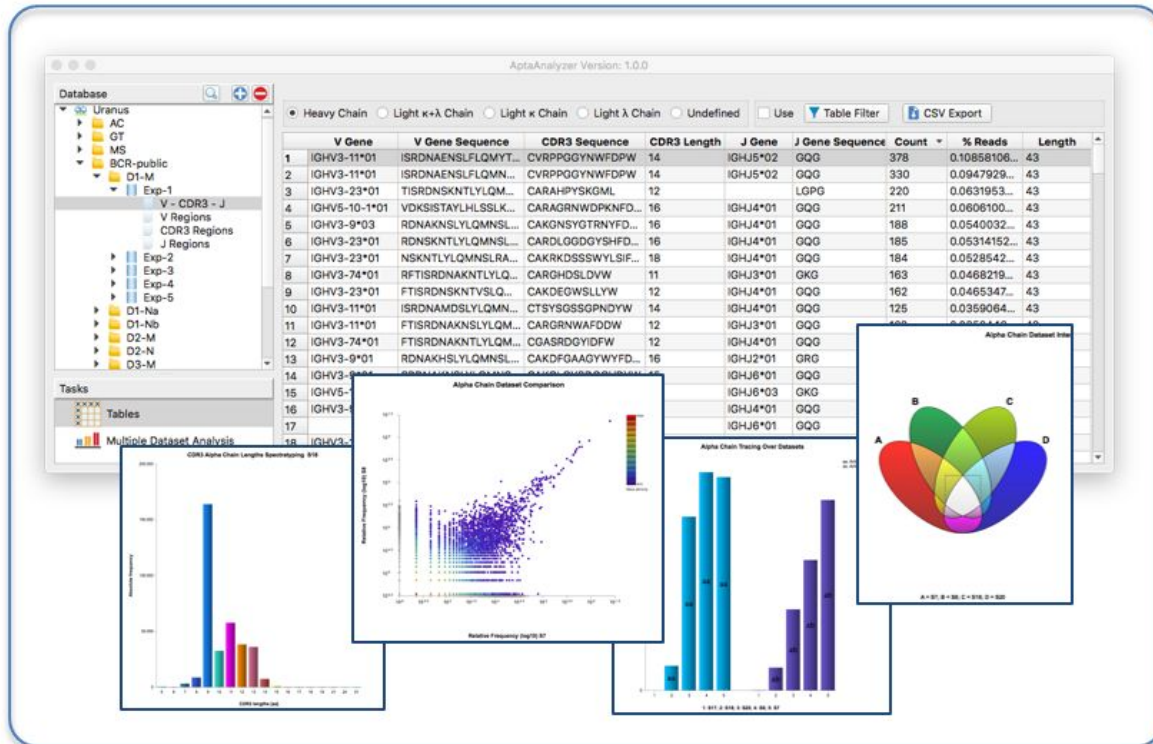
Overview

NGS Data Analysis: the basic idea

<http://bit.ly/2r1Y2Dr>

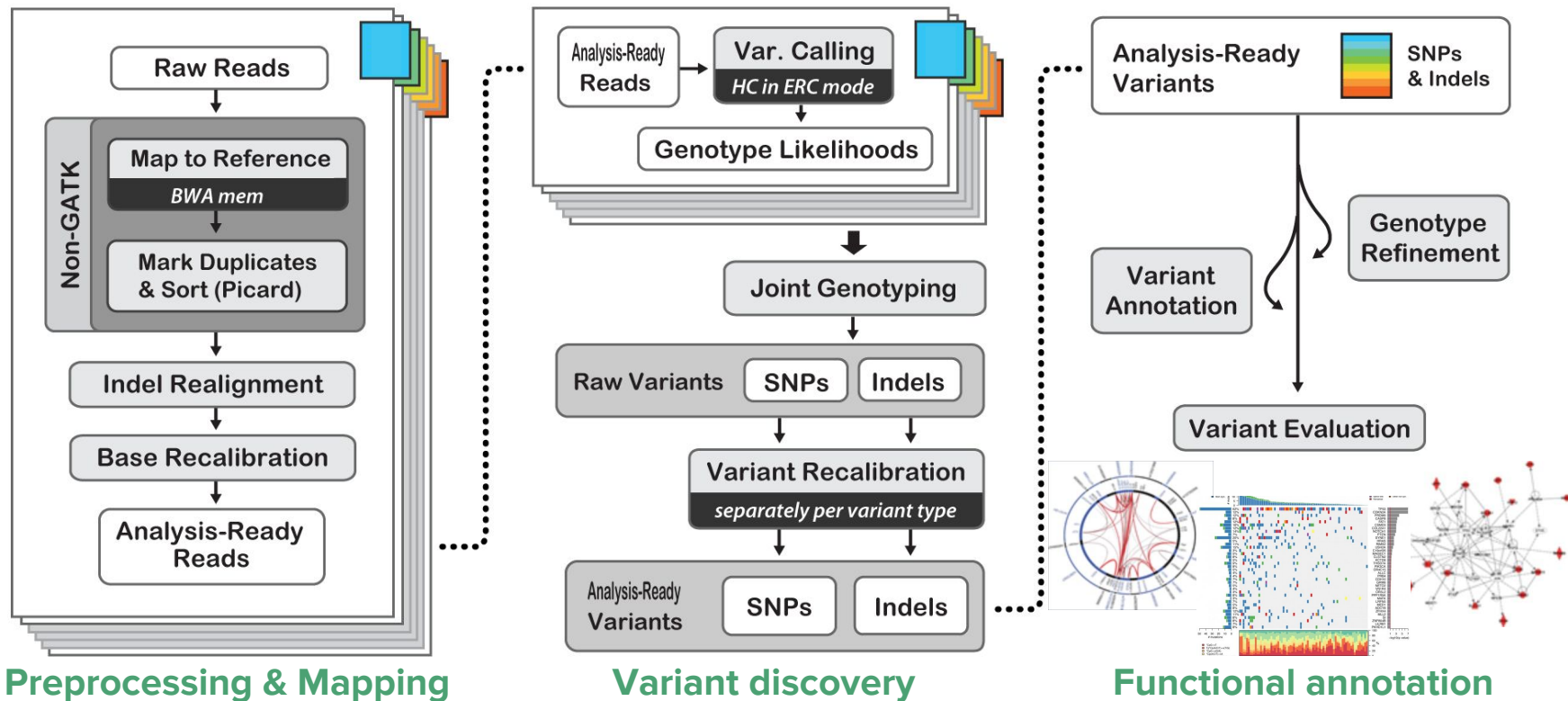


NGS data



Organization and analysis of NGS data

NGS Data Analysis: the actual workflow



NGS data analysis

Quality check & preprocessing

Fasta files

>J01415.2 Homo sapiens mitochondrion, complete genome

GATCACAGGTCTATCACCTATTAACCACTCACGGGAGCTCTCCATGCATTGTTTTCGTCTGGGGG
GTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCTATGTCGCAGTATCTGTCTTTGATTC
CTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACTTACTAAAGTGTGTTA
ATTAATTAATGCTTGTAGGACATAATAATAACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATC
ATAACAAAAAATTTCCACCAAACCCCCCTCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCA

Sequence ID

Sequence

- Both human- and machine-readable
- Can store multiple sequences
- ID can contain details or comments
- Usually contains full genomes or long sequence chunks

Fastq files

```
@HWUSI-EAS100R:6:73:941:1973#0/1 } Sequence ID
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCACAGTTT } Sequence
+ Spacer
!' '*((( (**+))%%++) (%%%) .1***-+*' '))**55CCF>>>>>CCCCC420 } Quality Score
```

- Both human- and machine-readable
- Can store multiple sequences
- ID can contain sequencing details and technical info
- Usually contains short sequence chunks (sequencing reads)

Quality score

Phred quality score: estimated probability of an error in base calling

usually [0-40]

$$Q_{\text{sanger}} = -10 \log_{10} P$$

Probability that the
base call is incorrect

Encoded using ASCII characters in fastq files:

Quality score	Probability of errors	ASCII encoding
0-9	1	!"#\$%&'()*
10-19	1/10	+,-./01234
20-29	1/100	56789:;<=>
30-39	1/1000	?@ABCDEFGH
40	1/10000	I

Quality score

@HWUSI-EAS100R:6:73:941:1973#0/1

GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

+

! ' * (((* * * +)) % % + +) (% % %) . 1 * * * - + * ' ')) * * 5 5 C C F > > > > > C C C C C C 4 2 0

0 6 6 9 34 19 17 15

Quality score	Probability of errors	ASCII encoding
0-9	1	!"#\$%&'()*
10-19	1/10	+,-./01234
20-29	1/100	56789:;<=>
30-39	1/1000	?@ABCDEFGH
40	1/10000	I

Quality check

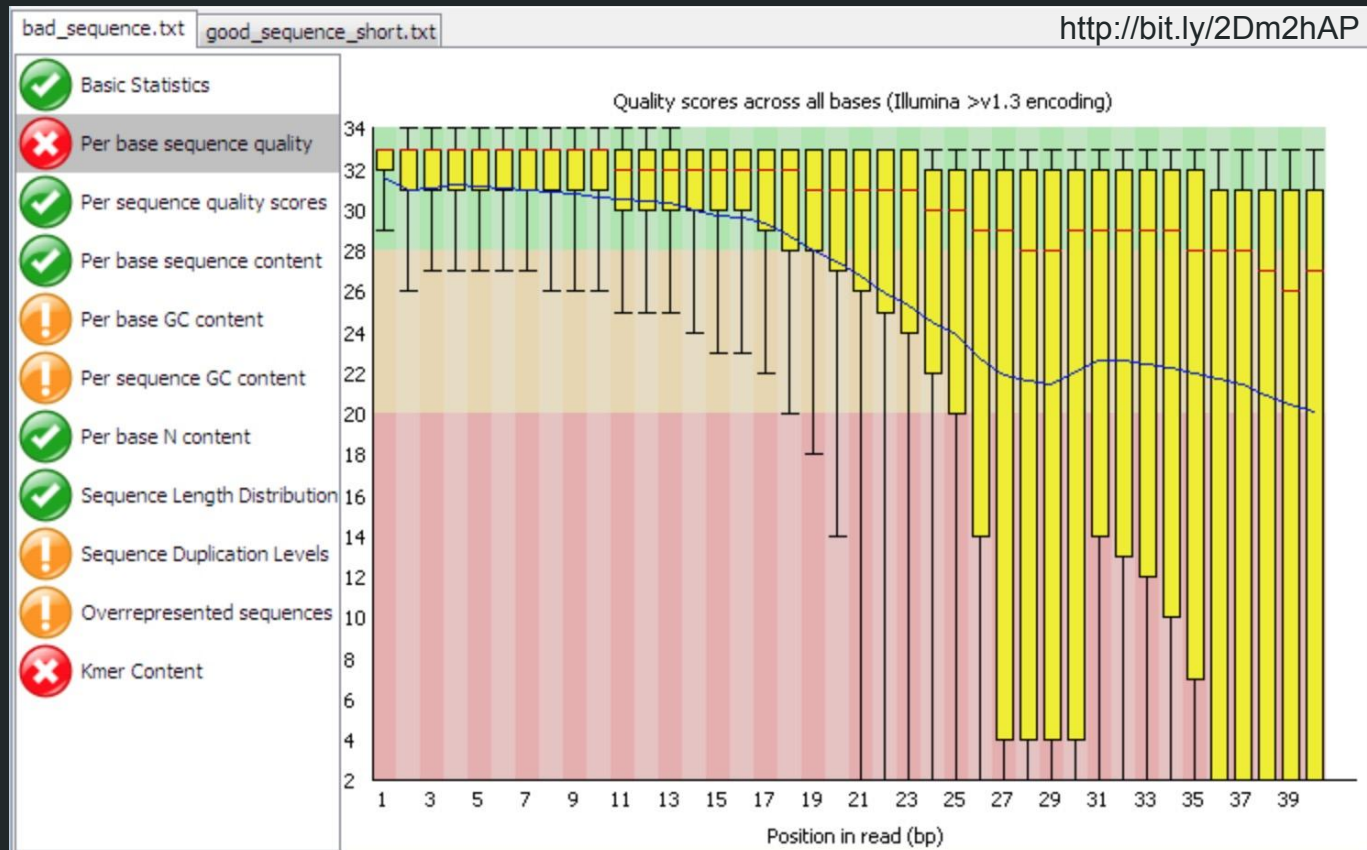
FastQC: visual report of several quality checks for NGS data, useful for further processing

Different
modules
=
different
checks

Pass

Warning

Fail



FastQC modules

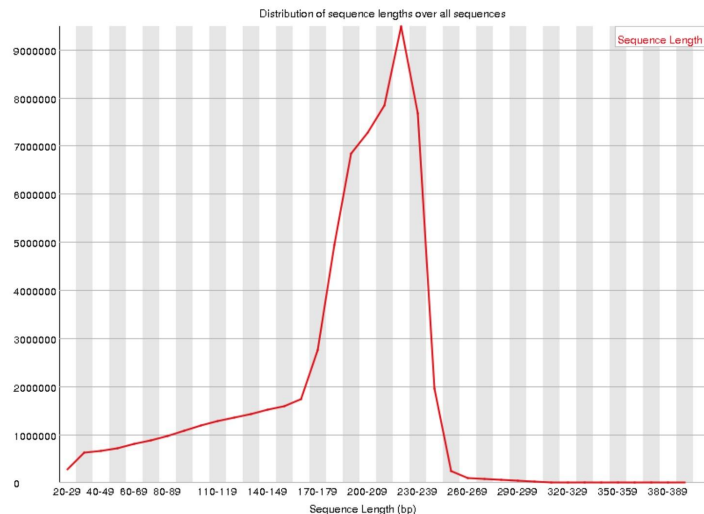


Basic Statistics

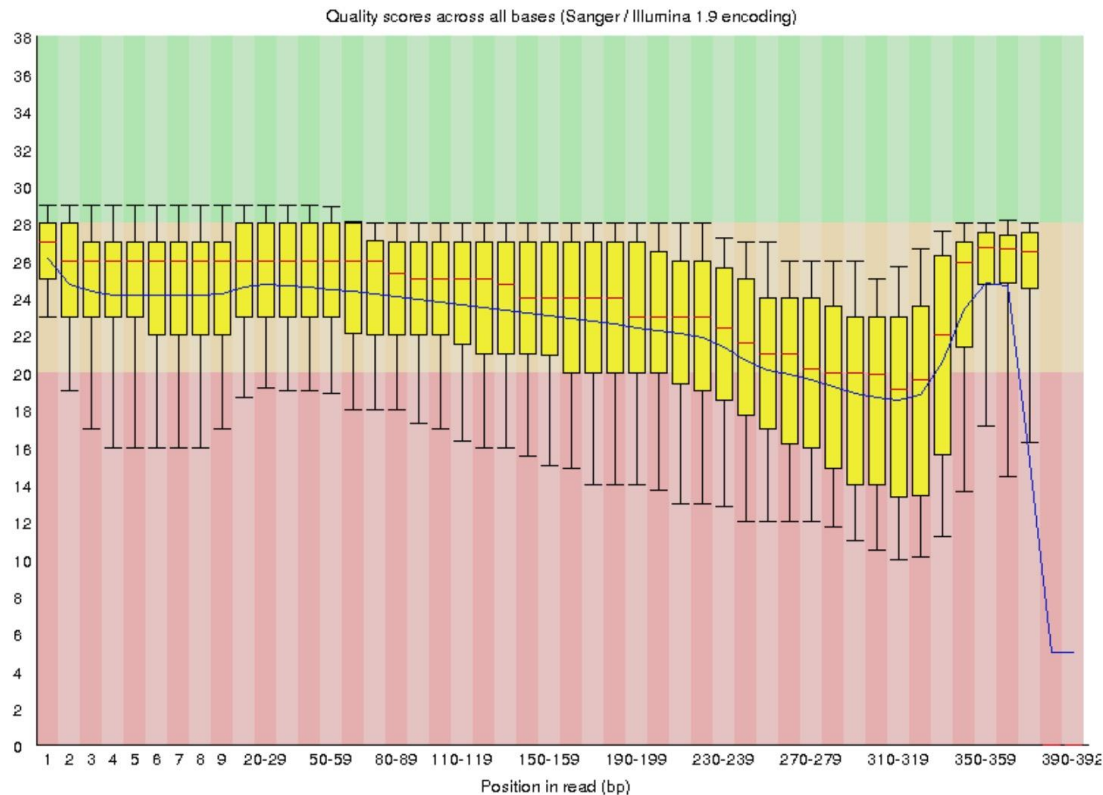
Measure	Value
Filename	ionXpress002.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	65578765
Sequences flagged as poor quality	0
Sequence length	25-392
%GC	52



Sequence Length Distribution



Per base sequence quality



Preprocessing

Cleansing of reads to solve several issues:

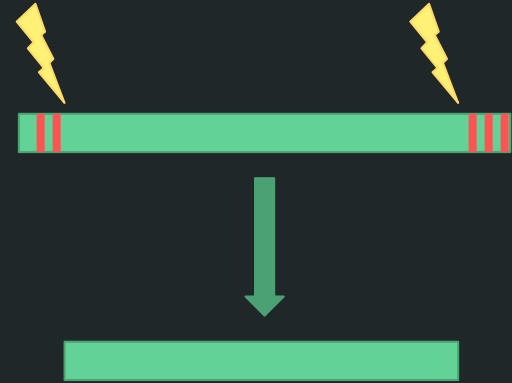
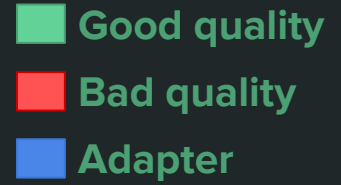
- remove adapters



Preprocessing

Cleansing of reads to solve several issues:

- remove adapters
- cut low quality bases from both ends



Preprocessing

Cleansing of reads to solve several issues:

- remove adapters
- cut low quality bases from both ends
- drop short reads

■ Good quality
■ Bad quality
■ Adapter



Preprocessing

Cleansing of reads to solve several issues:

- remove adapters
- cut low quality bases from both ends
- drop short reads

Common tools:

- Trimmomatic
- Trim Galore!
- FASTX

Post-processing quality check

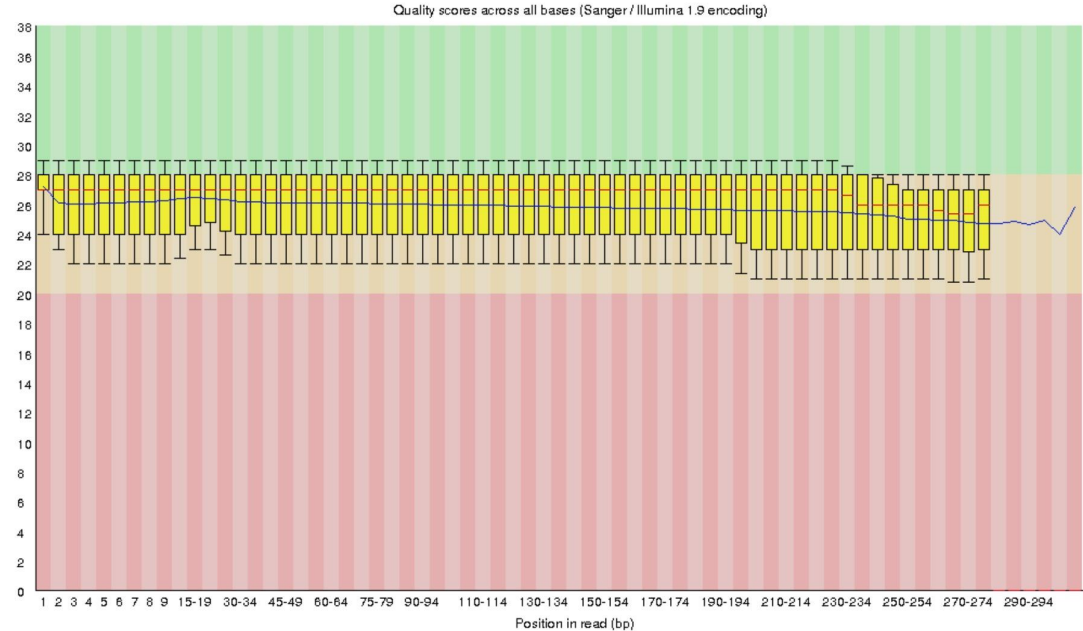


Basic Statistics

Measure	Value
Filename	ionXpress002.trimmed.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	34710199
Sequences flagged as poor quality	0
Sequence length	30-307
%GC	53



Per base sequence quality



Was the processing effective?

Are these data ready to be aligned?

NGS data analysis

Alignment

Alignment vs Assembly

Alignment (*reference-based*)

reference genome available

→ reads aligned on it

..GTGACTTAGTCGTAGCTAGCTAGCTAGCTCGATCTAGA..



GTGACTTAGT

TAGTAGCTCG

GAGTTAGTCG

GTAGCTCGAT

CTTAGTCGTA

AGCTCGACCT

TAGTCGTAGC

CTCGACCTAG

Assembly (*de-novo*)

reference genome not available

→ reads aligned with each other

GTGACTTAGT

GCTAGCTAGT

CGATCTAGA

AGTTAGTCGT

GTAGCTCGAT

TCGTAGCTAG

AGCTCGACCT

TGAGTTAGCC

AGCTAGTAGC



..GTGACTTAGTCGTAGCTAGCTAGCTAGCTCGATCTAGA..

Alignment

..GTGACTTAGTCGTAGCTAGCTAGCTAGCTCGATCTAGA..

Reference genome

GTGACTTAGT

TAGTAGCTCG

GAGTTAGTCG

GTAGCTCGAT

CTTAGTCGTA

AGCTCGACCT

TAGTCGTAGC

CTCGACCTAG

Reads

Common aligners:

- BWA
- Bowtie
- GMAP/GSNAP



<http://bit.ly/2DG8lQj>

Assembly

<http://bit.ly/2DnwRdt>

GTGACTTAGT GCTAGCTAGT CGATCTAGA
AGTTAGTCGT GTAGCTCGAT
 TCGTAGCTAG AGCTCGACCT
TGAGTTAGCC AGCTAGTAGC

Reads



..GTGACTTAGTCGTAGCTAGCTAGTAGCTCGATCTAGA..

Consensus sequence

Common approaches:

- greedy algorithm
- graph method

Common assemblers:

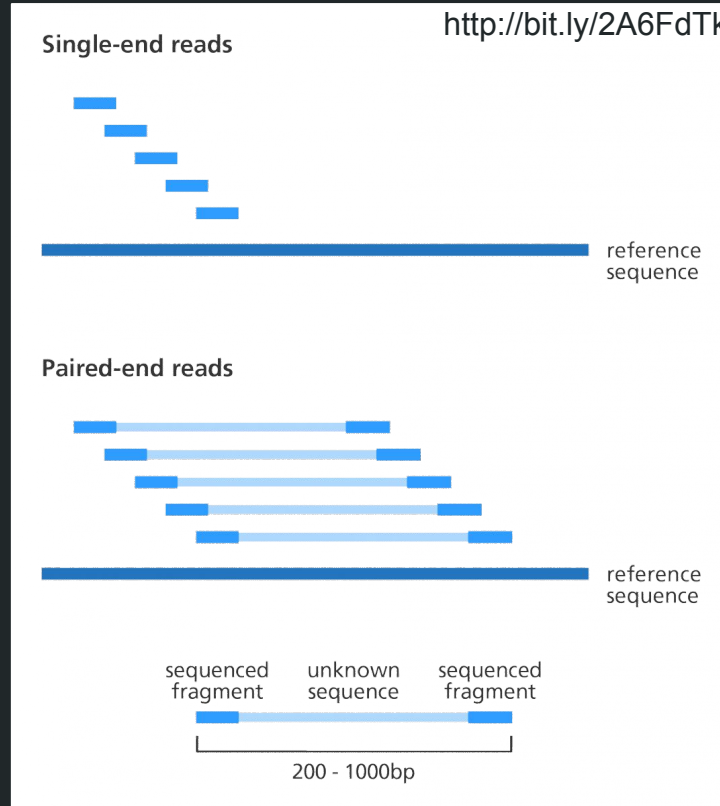
- Newbler
- SPAdes
- MaSuRCA

Paired-end vs single-end reads

Easy to identify read position
in genome

High accuracy for structural
rearrangements and
assembly of repetitive
regions

More expensive



Suitable for most applications

Cheaper

Faster

SAM/BAM files

Text-based format used to store aligned reads (to a reference genome)

Header {

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

} Alignments

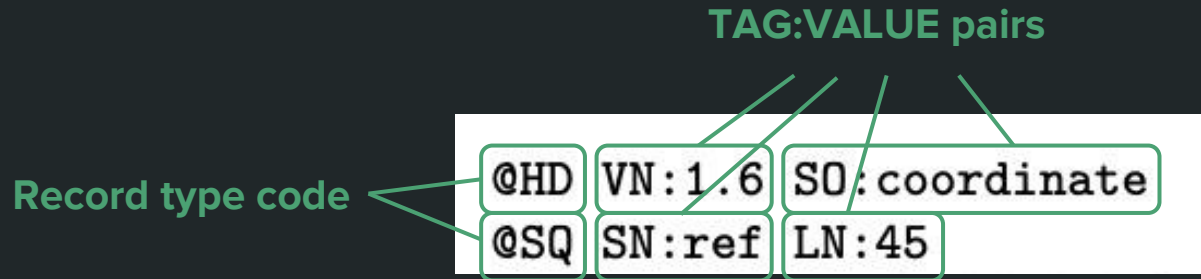
SAM (Sequence Alignment Map)

- Both human- and machine-readable
- Header section contains TAG:VALUE pairs
- Alignment section contains 11 mandatory fields

BAM (Binary Alignment Map)

- Compressed version of SAM
- Binary format
- Only machine-readable

SAM files



@HD: header line delimiter

@SQ: reference sequence

VN: version number

SN: reference sequence name

SO: alignments sorting order

LN: reference sequence length

SAM files

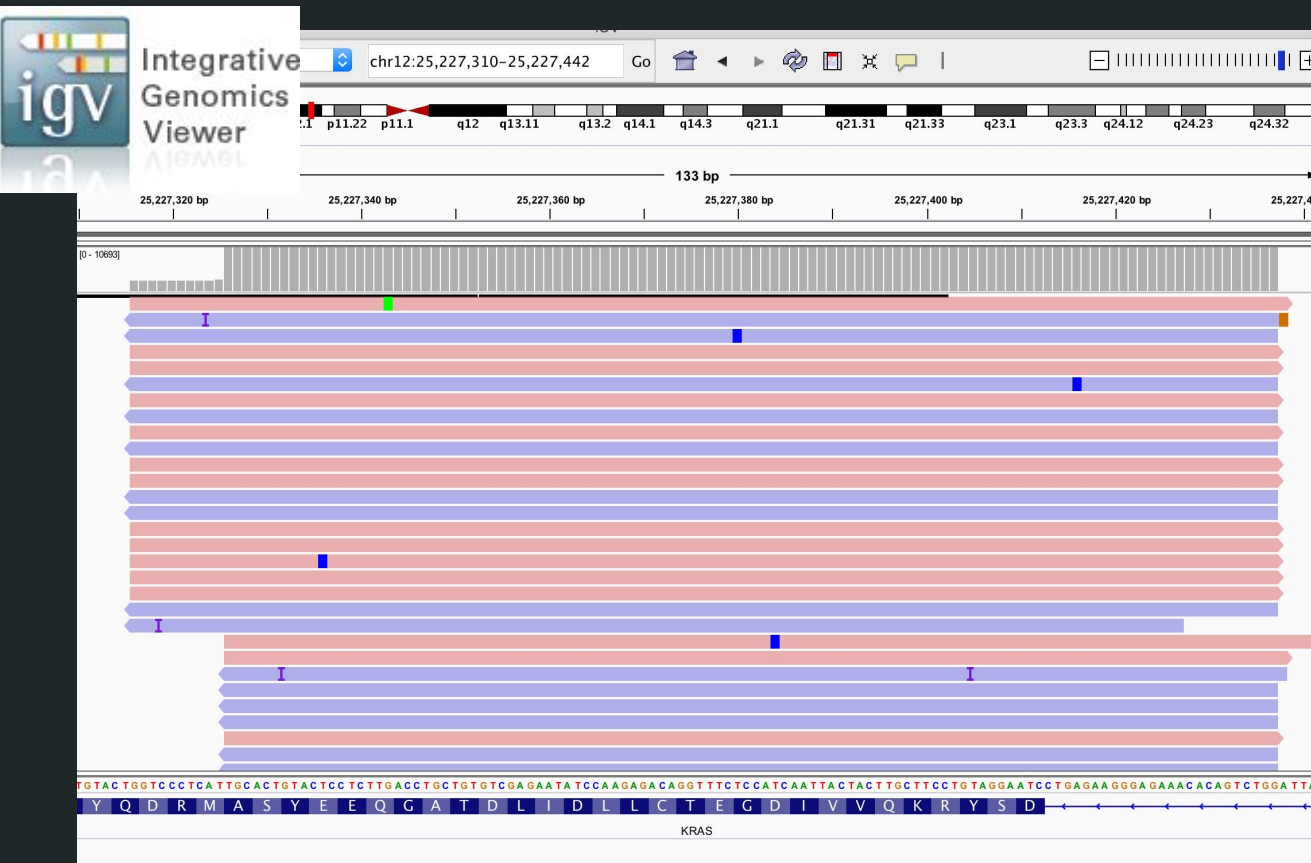
```

r001    99 ref    7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002     0 ref    9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003     0 ref    9 30 5S6M          * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004     0 ref   16 30 6M14N5M      * 0 0 ATAGCTTCAGC *
r003 2064 ref   29 17 6H5M          * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001   147 ref   37 30 9M           = 7 -39 CAGCGGCAT * NM:i:1
  
```

POS CIGAR

Col	Field	Type	Regex/Range	Brief description
1	QNAME	String	[!~?A~]{1,254}	Query template NAME
2	FLAG	Int	[0, 2 ¹⁶ - 1]	bitwise FLAG
3	RNAME	String	* [!-()+-<>~] [!~]*	Reference sequence NAME
4	POS	Int	[0, 2 ³¹ - 1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0, 2 ⁸ - 1]	MAPping Quality
6	CIGAR	String	* ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	* = [!-()+-<>~] [!~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0, 2 ³¹ - 1]	Position of the mate/next read
9	TLEN	Int	[-2 ³¹ + 1, 2 ³¹ - 1]	observed Template LENgth
10	SEQ	String	* [A-Za-z=.]+	segment SEQUENCE
11	QUAL	String	[!~]+	ASCII of Phred-scaled base QUALity+33

Alignment quality check



Integrative Genomics Viewer (IGV)

Interactive visualization of NGS data from Fasta, SAM, BAM, VCF files

Additional features are organized in tracks (gene expression, methylation, copy number variations...)

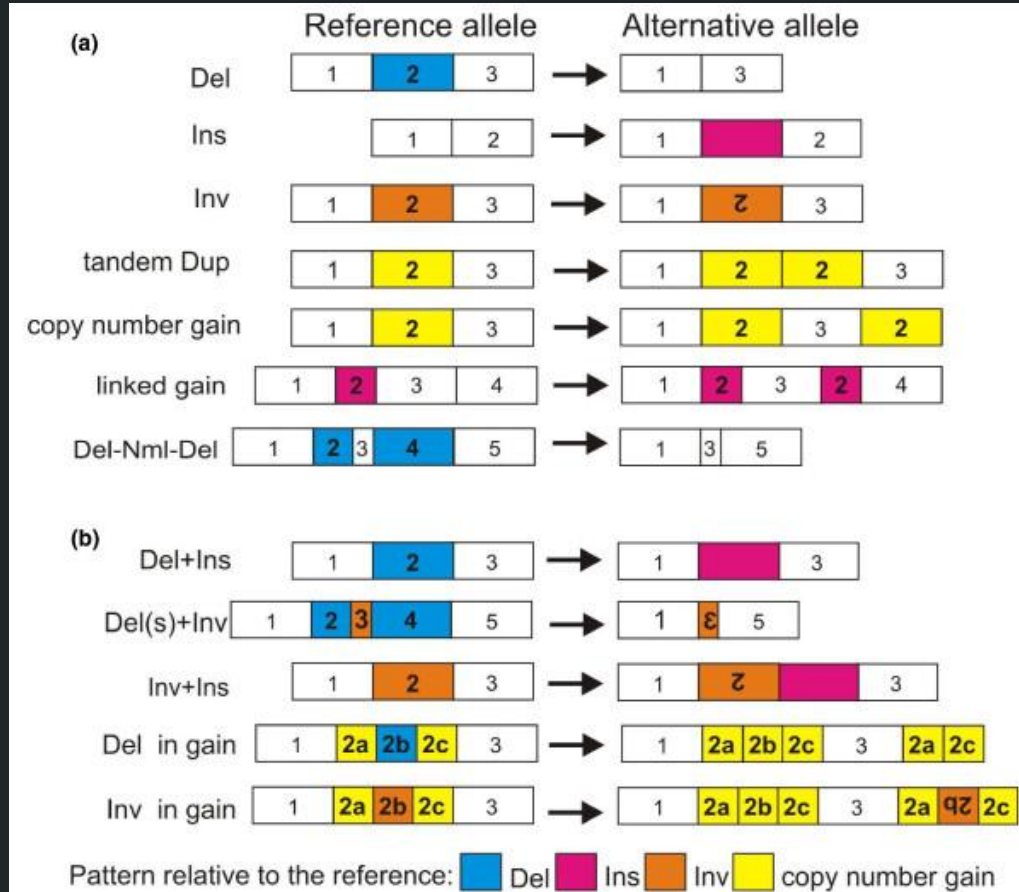
<http://bit.ly/2qNbvP0>

NGS data analysis

Variant calling

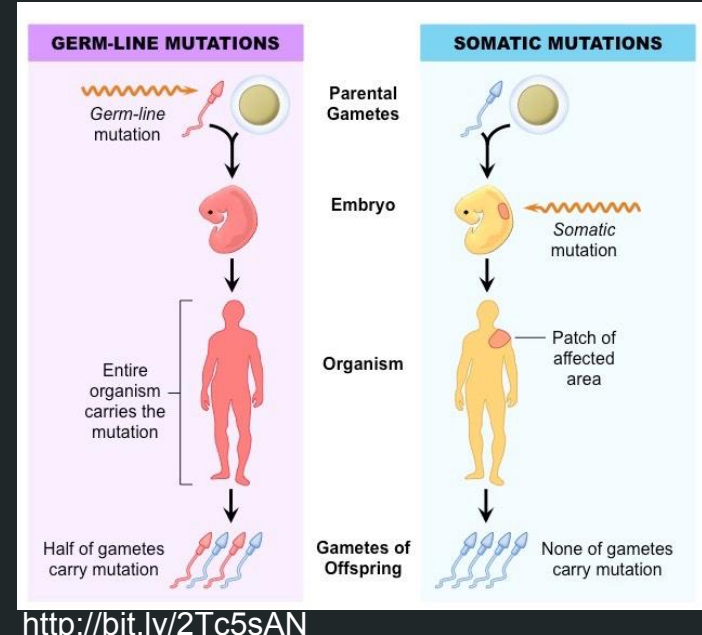
Variations

<http://bit.ly/2QHQocd>



Aligned data can be used to assess the presence of mutations:

- somatic
- germline

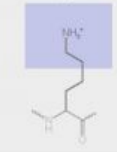
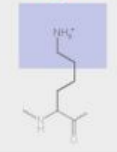
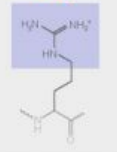
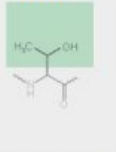


Variations

Most common variations searched for:

	Variation type
Reference	ACTGACGCATGCATCATGCATGC
SNP	ACTGACGCATGCATCAT T CATGC
Insertion	ACTGACGCATG GT ACATCATGCATGC
Deletion	ACTGACGC -- GCATCATGCATGC

Variation effect

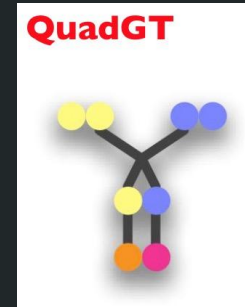
	No mutation	Point mutations			
		Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	Lys	Lys	STOP	Arg	Thr
					

<http://bit.ly/2K9a3zj>

Variant callers

A plethora of different tools, each with its own peculiarities:

- variation type (SNV vs structural variation)
- source (somatic vs cancer mutations)
- only detect variants vs also predict their effect



VCF files

<http://bit.ly/2Kfl80W>

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

Header
(information and metadata)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0 0:48:1:51,51	1 0:48:8:51,51
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0 0:49:3:58,50	0 1:3:5:65,3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1 2:21:6:23,27	2 1:2:0:18,2
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0 0:54:7:56,60	0 0:48:4:51,51
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2

Variants

**Variant
annotations**

**Genotype
annotations**

VCF files

Mandatory fields:

#CHROM: chromosome or contig

POS: variant position (1-based)

ID: variant identifier (usually dbSNP ID)

REF: reference base(s)

ALT: alternative base(s)

QUAL: Phred quality score for each ALT

FILTER: variant call filter status

INFO: key-value pairs with additional information (described in header)

Optional fields:

FORMAT: genotype information fields

SAMPLE1 ... SAMPLEn: values for fields listed in FORMAT

NGS data analysis

Functional annotation

Functional annotation

~3.000.000 small variants
(SNP/indels) per genome

ATCATGCATGC

ATCATTCATGC



<http://bit.ly/2QekbN3>



<http://bit.ly/2BfefuO>

Which ones are most
interesting for our
purpose?



<http://bit.ly/2KfpRRt>

Annovar

Identify variants and flag those with detrimental effects

ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data

Kai Wang^{1,*}, Mingyao Li² and Hakon Hakonarson^{1,3}

<http://bit.ly/2zc9fp5>

Gene-based annotations

- identify variants that can cause protein coding changes
- detect the amino acids that are affected

Annovar

Identify variants and flag those with detrimental effects

ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data

Kai Wang^{1,*}, Mingyao Li² and Hakon Hakonarson^{1,3}

<http://bit.ly/2zc9fp5>

Region-based annotations

→ identify variants in specific genomic regions:

- conserved regions
- predicted transcription factor binding sites
- segmental duplication regions
- etc

AnnoVar

Identify variants and flag those with detrimental effects

ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data

Kai Wang^{1,*}, Mingyao Li² and Hakon Hakonarson^{1,3}

<http://bit.ly/2zc9fp5>

Filter-based annotations

→ identify variants that are documented in specific databases:

- dbSNP
- 1000 Genome Project
- etc

Annovar VCF files

Extensive number of new annotations added to the initial VCF file

Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	GeneDetail.refGene	ExonicFunc.refGene	AAChange.refGene	Xref.refGene	ExAC_Freq
1	948921	948921	T	C	UTR5	ISG15	NM_005101:c.-33T>C	.	.	Immunodeficiency	0.941
1	1404001	1404001	G	T	UTR3	ATAD3C	NM_001039211:c.*91G>T	.	.	.	0.054
1	5935162	5935162	A	T	splicing	NPHP4	NM_001291594:exon17:c.1282-2T>A	.	.	Nephronophthosis	0.823
1	162736463	162736463	C	T	intronic	DDR2	.	.	.	Spondylometaphyseal dysplasia	.
1	84875173	84875173	C	T	intronic	DNASE2B
1	13211293	13211294	TC	-	intergenic	PRAMEF36P;F	dist=11566;dist=116902
1	11403596	11403596	-	AT	intergenic	UBIAD1;PTCH	dist=55105;dist=135699
1	105492231	105492231	A	ATAAA	intergenic	LOC1001291:	dist=872538;dist=640085
1	67705958	67705958	G	A	exonic	IL23R	.	nonsynonymous SNV	IL23R:NM_144701:exon9:c.G1142A:p.R381Q	.	0.041
2	234183368	234183368	A	G	exonic	ATG16L1	.	nonsynonymous SNV	ATG16L1:NM_198890:exon5:c.A409G:p.T137A;ATG16L1:NM_001101:exon5:c.A409G:p.T137A	.	0.457
16	50745926	50745926	C	T	exonic	NOD2	.	nonsynonymous SNV	NOD2:NM_001293557:exon3:c.C2023T:p.R675W;NOD2:NM_001293557:exon3:c.C2023T:p.R675W	Blau syndrome, A	0.023
16	50756540	50756540	G	C	exonic	NOD2	.	nonsynonymous SNV	NOD2:NM_001293557:exon7:c.G2641C:p.G881R;NOD2:NM_001293557:exon7:c.G2641C:p.G881R	Blau syndrome, A	0.009917
16	50763778	50763778	-	C	exonic	NOD2	.	frameshift insertion	NOD2:NM_001293557:exon10:c.2936dupC:p.L980Pfs*2;NOD2:NM_001293557:exon10:c.2936dupC:p.L980Pfs*2	Blau syndrome, A	0.013
13	20763686	20763686	G	-	exonic	GJB2	.	frameshift deletion	GJB2:NM_004004:exon2:c.35delG:p.G12Vfs*2	Bart-Pumphrey syndrome	0.006038
13	20797176	21105944	O	-	exonic	CRYL1;GJB6	.	frameshift deletion	GJB6:NM_001110220:wholegene;GJB6:NM_001110221:wholegene	.	.
8	8887543	8887543	A	T	exonic	ERI1	.	stoploss	ERI1:NM_153332:exon7:c.A1049T:p.X350L	.	.
8	8887539	8887539	A	T	exonic	ERI1	.	stopgain	ERI1:NM_153332:exon7:c.A1045T:p.K349X	.	.
8	8887536	8887537	AG	GATT	exonic	ERI1	.	frameshift substitution	ERI1:NM_153332:exon7:c.1042_1043GATT:p.R348Dfs*2	.	.
8	8887540	8887540	G	GGAA	exonic	ERI1	.	nonframeshift substitution	ERI1:NM_153332:exon7:c.1046delinsGGAA:p.R348_K349insR	.	.
5	1295288	1295288	G	A	upstream	TERT	dist=126

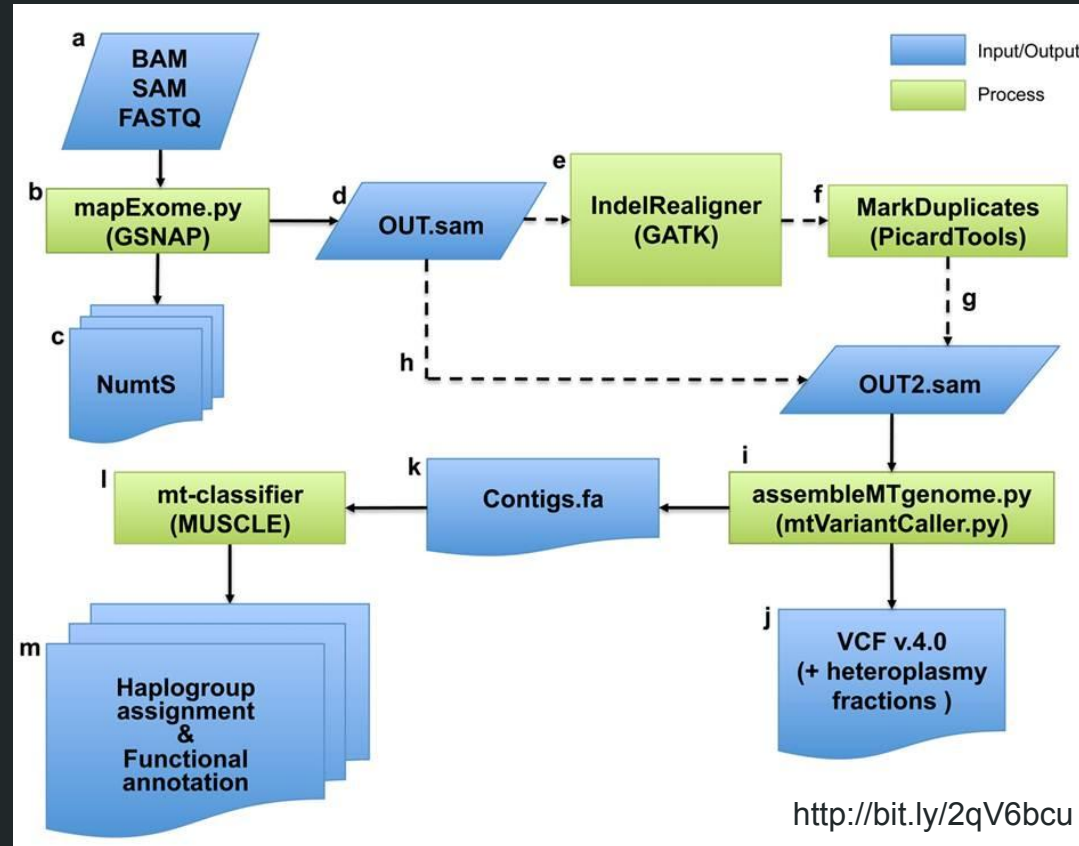
<http://bit.ly/2Q6cvfr>

MToolBox

Human mtDNA reconstruction,
analysis and annotation from NGS
data

Haplogroup predictions

Both command-line and
web-based versions available

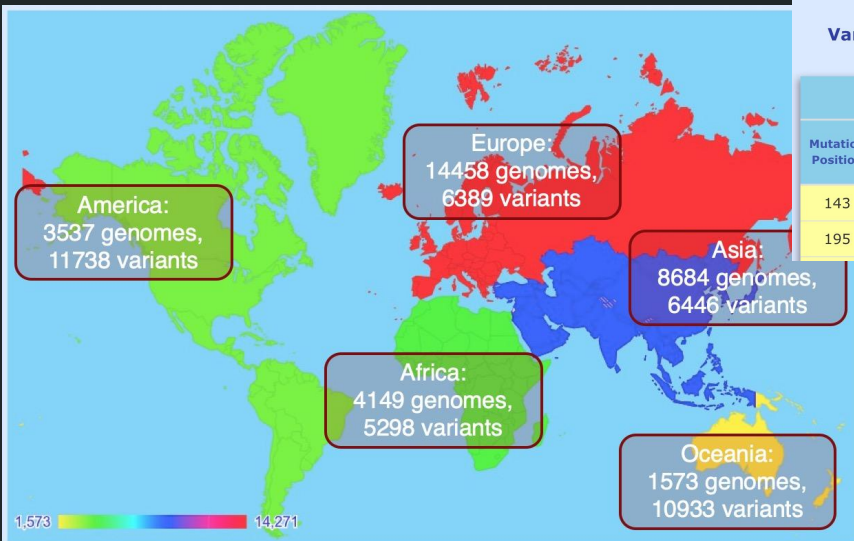


HmtDB

Over 48000 human mitochondrial sequences

Healthy/patient and continent-specific subsets

Genome-centric



HmtDB Genome Card

Identifier: AM_BS_0001

Reference DB: Genbank

Reference DB Source Id: **FJ769771**

HmtDB Assigned Haplogroup: L2a111a

Author Assigned Haplogroup: L2a1

Haplotype User Code: Undefined

Genome Length: 16568 - complete genome

Source: Undefined

Sequencing Method: Undefined

References: Unpublished

Individual's Data

Continent	Country	Ethnic Group	Age	Sex	Phenotype
America	Bahamas	Black American	Undefined	Undefined	Normal

Variants' Data

synonymous

non-synonymous

d-loop

rna

non coding

Mutations vs RSRS					Site-specific Variability Data					
Mutation Position	Mutation Type	AA Position	AA Change	Locus Name	Human NT Site Variability Normal	Human NT Site Variability Patient	Human AA Site Variability Normal	Human AA Site Variability Patient	AA Variability InterMammals	Disease Associations (Mitomap)
143	G → A			DLOOP	0.065	0.023				
195	C → T			DLOOP	0.539	0.345				



<https://www.hmtdb.uniba.it>

HmtVar

Over 40000 human mitochondrial variants

Pathogenicity predictions available for most variants

Variant-centric

HmtVar

The main web resource to explore human mitochondrial variability data and their pathological correlation.

[Query HmtVar](#) [HmtVar API](#)

Latest update: **September 2018.**

Variant Card

Variability and pathogenicity data available for T3440G.

[Main Info](#) [Variability](#) [Pathogenicity Predictions](#) [External Resources](#) [Download Data](#)

Main Info

Main information regarding the **T3440G** variant.

Basic		
Position: 3440	Locus Type: Coding Sequence	HmtVar Prediction: Pathogenic
Mutation: T → G	Locus: MT-ND1	More Info
Codon position: 2		
Aa Change: Leu → Arg, aa position 45		

<https://www.hmtvar.uniba.it>

NGS Data Analysis: pipelines



General-purpose programming language

Easy to learn

Very powerful (libraries available for anything you can think of)

→ Biopython: specific module for bioinformatics

NGS Data Analysis: pipelines

Originally suited for statistics

Particularly used for data analysis and visualization

Gained a lot of traction for many different disciplines

→ Bioconductor: specific package for bioinformatics



Useful info

Contacts:

roberto.preste@gmail.com

Slides:

<http://bit.ly/ngs-data>