Genomic variants

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

3/4/2021

Midterm

10:30am – 1pm on 3/11

Text editor
Unix
NGS technology
R
NGS tools
alignment (principle and Blast)
alignment II

Homework 4 will help you answer exam questions *review*

Alignment algorithms

Genome sequences

“sorted or indexed” genome

BWT

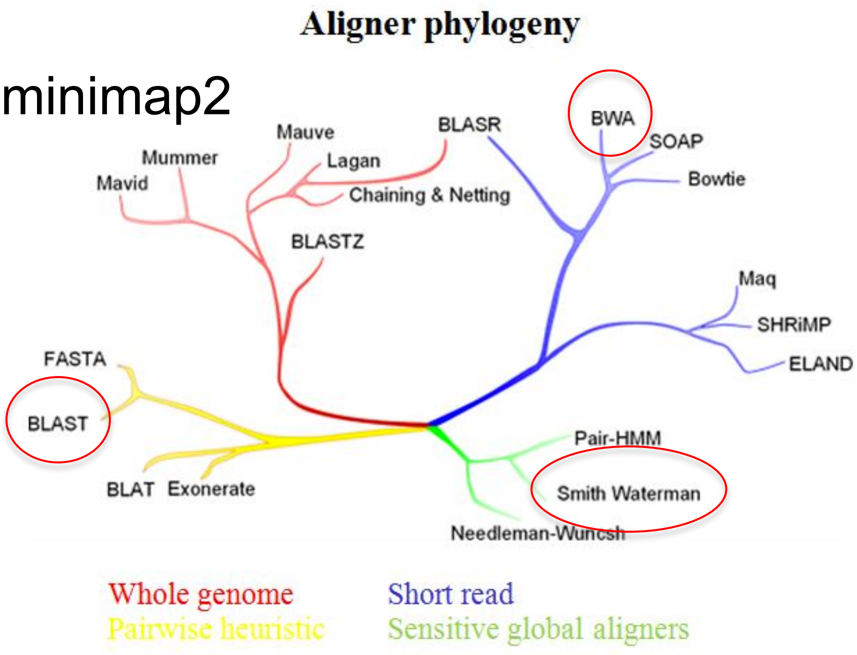
reads

one by one

alignments

minimap2

Aligner phylogeny



evomicsorg.wpengine.netdna-cdn.com/wp-content/uploads/2014/01/alignCompare2.jpg

bwa index

bwa mem

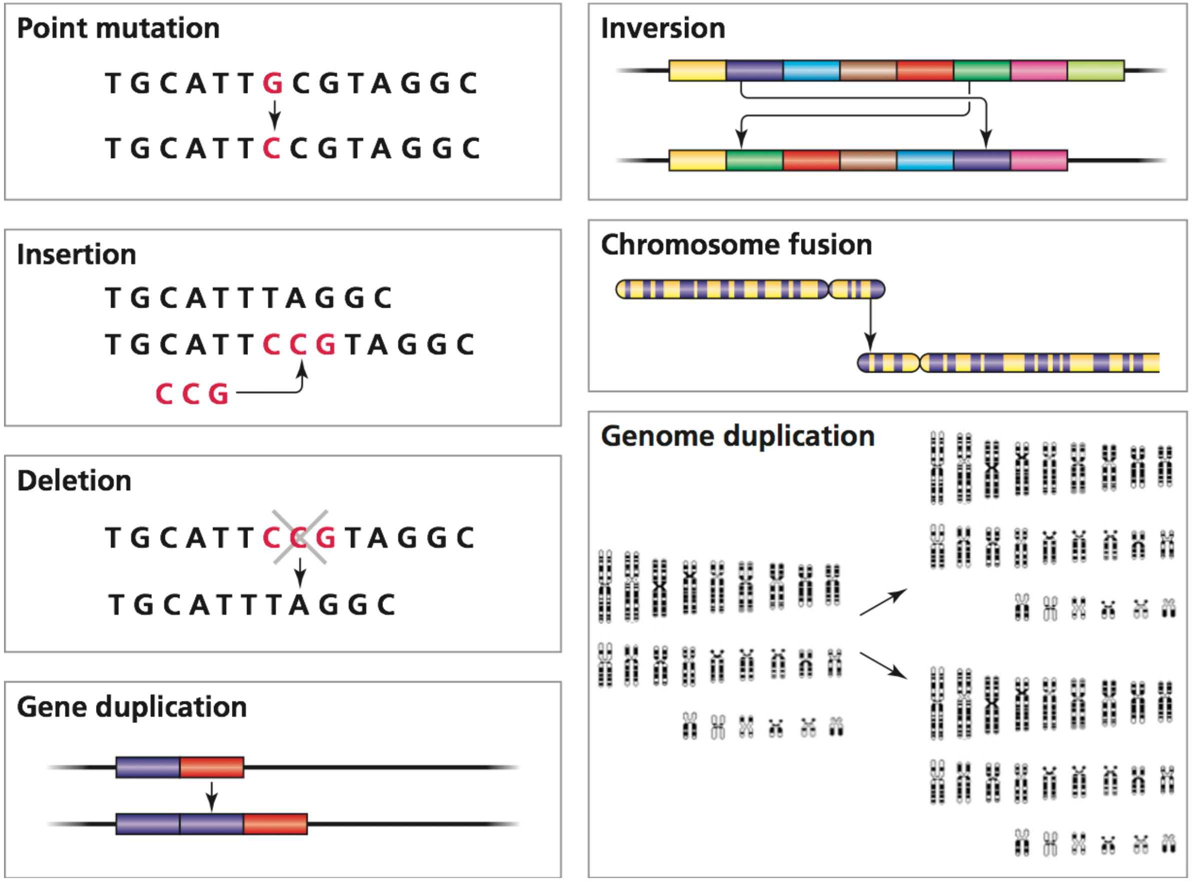
Outline

- Overview of genomic variants
- Data for variant discovery
- Bioinformatics of variant discovery
- the methods for variant (SNP) validation

Genomic variants (ploymorphisms)

1. SNP

2. INDEL



3. genomic structural variation

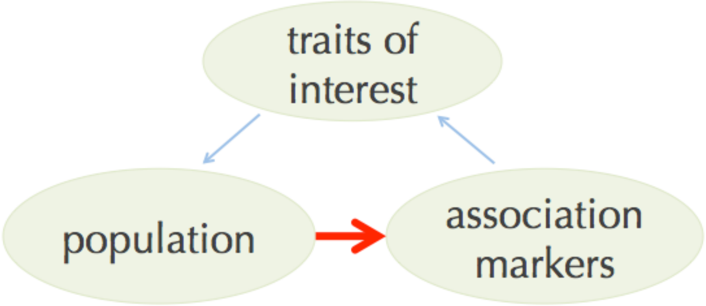
- copy number variation (presence-absence variation)
- other re-arrangements

Genomic variants - SNPs

- SNP stands for single nucleotide polymorphism.
- Frequencies of SNPs are depended on species. For example, millions of SNPs have been discovered in human.
- Most SNPs are bi-allelic. (mutation rate per site is about 10^{-8})
- Most have no functional effects but some could have important phenotypic consequences.

Applications of SNPs

1. Genetic markers to map the genetic controlling of traits (quality traits, quantitative traits, gene expression, etc)

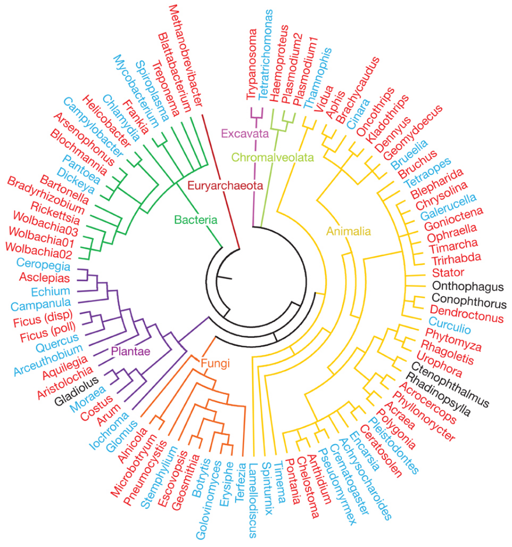


traits of interest

population

association markers

2. Genetic markers to construct genetic maps
3. Markers to construct phylogenetic trees



To monitor pathogen evolution

Next-Generation Sequencing to generate data for variant discovery

GATCTGCGTCATACGGAAT

GATCTGCGTGATACGGAAT

GATCTGCGTCATACGGAAT

GATCTGCGTGATACGGAAT

GATCTGCGTGATACGGAAT

GATCTGCGTGATACGGAAT

GATCTGCGTCATACGGAAT

GATCTGCGTCATACGGAAT

GATCTGCGTGATACGGAAT

GATCTGCGTCATACGGAAT

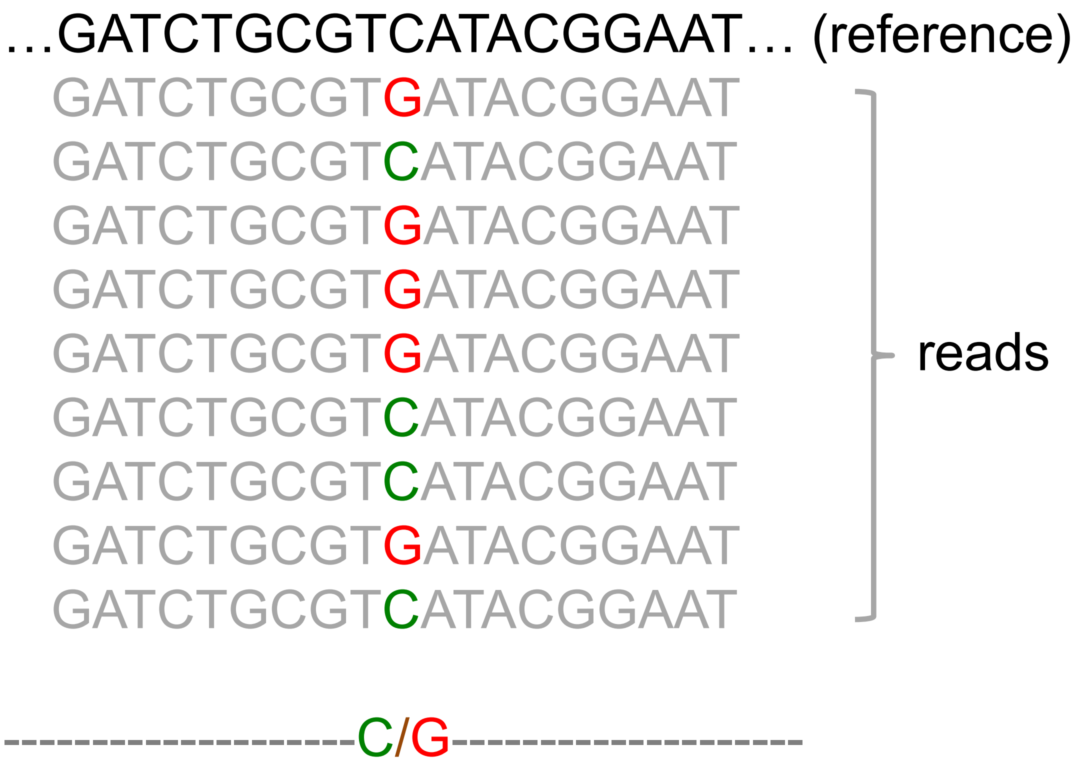
---------C/G---------

Heterozygous call
(diploid genome)

Approaches for data generation

- **Whole genome sequencing** (WGS): high genome coverage but costly for large genomes
- **Exome-capture sequencing**: target on genic regions but still expensive to perform large number of samples
- **RNA sequencing** (RNA-Seq): obtain data on genic regions and provide expression information
- **Genotyping-By-Sequencing** (GBS): cost-efficient and high-throughput approach

Alignment-based SNP discovery



```
...GATCTGCGTCATACGGAAT... (reference)
   GATCTGCGTGATACGGAAT
   GATCTGCGTCATACGGAAT
   GATCTGCGTGATACGGAAT
   GATCTGCGTGATACGGAAT
   GATCTGCGTGATACGGAAT         reads
   GATCTGCGTCATACGGAAT
   GATCTGCGTCATACGGAAT
   GATCTGCGTGATACGGAAT
   GATCTGCGTCATACGGAAT

-------------------C/G-------------------
```

Alignment-based SNP discovery, cont.

General procedure

- Reads cleanup (adaptor, quality trimming, e.g., trimmomatic)
- Reads aligned to the reference genome with aligners
  1. BWA, Bowtie (DNA-Seq reads)
  2. HISAT2, STAR, GSNAP, Tophat (RNA-Seq reads)
- Post-alignment filtering and convert SAM (alignment file) to BAM
- SNP calling with software packages: Samtools, GATK, VarScan2
- Use population information or some criteria to filter SNP sets

an example of the alignment of a RNA-Seq read using GSNAP

```
HISEQ:163:C4YWTACXX:1:1101:8654:2286  16    chromosome_1     1765703    40    16M97N85M  *     0     0
TGGCAACATTGTTAAGCTCTGGCGTGATAAATTGCCCTGTCAGCAAGCAGATGGCAGGCAATGTGCAATAGGCGAAGAGCGGGATAGATGTCCAAGGGTA
T
DDDDDDDDDDDDDDDDDDDBDDCEEEFEFFFFFHHHJJIJJJJJJJJJJJJJJJJJJJJJJJJJIJJJJJHEJJJJJJJJJJJJJHHHHHFFFFFCC
C     MD:Z:73C27 NH:i:1     HI:i:1     NM:i:1     SM:i:40    XQ:i:40    X2:i:0     XO:Z:UU    XS:A:-
      PG:Z:A
```

Interpretation of the BWA alignment

SAM output:

```
HWI-ST897:104:C015GACXX:6:1101:12678:20443  163 U00096
1888286 60  64M1D20M    =   1888358 170
GCCAACAGCCGCGACTTCCTGTACGCCAGGATGCTGCATGACGACATCTTCAATCTCGTTGGGAAGACGTTAAAAACGGAAACC   CCCFFFFFHHFHHJJJJJJJ
JHIJHIJIIJIJJJJJIJJJJIJJHHFFFFFFEEEEEEDDDDDDA5,53,8<?CC(50?8BD3?   NM:i:2  AS:i:72
XS:i:0  RG:Z:S1
```

CIGAR: 64M1D20M

NM: edit distance

edit distance is a way to quantify the dissimilarity of two strings (e.g., words) by counting the minimum number of edits (substitution, insertion, and deletion) required to transform one string into the other.

fact -> fit (2)

AACCT -> AAACT (1)

edit distance

- AACCT -> ACCTA (?)
- AATCCT -> ATCAT (?)

Polymorphism based on Alignment + reference genome

SAM output (BWA):

```
HWI-ST897:104:C015GACXX:6:1101:12678:20443  163 U00096
1888286 60  64M1D20M    =    1888358 170
GCCAACAGCCGCGACTTCCTGTACGCCAGGATGCTGCATGACGACATCTTCAATCTCGTTGGGAAGACGTTAAAAACGGAAACC    CCCFFFFFHHFHHJJJJJJJ
JHIJHIJIIJIJJJJJIJJJJIJJHHFFFFFFEEEEEEDDDDDDA5,53,8<?CC(50?8BD3?    NM:i:2  AS:i:72
XS:i:0  RG:Z:S1
```

mapping position and CIGAR determine the alignment

```
Query   1        GCCAACAGCCGCGACTTCCTGTACGCCAGGATGCTGCATGACGACATCTTCAATCTCGTT   60
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct   1888286  GCCAACAGCCGCGACTTCCTGTACGCCAGGATGCTGCATGACGACATCTTCAATCTCGTT   1888345

Query   61       GGGA-AGACGTTAAAAACGGAAACC   84
                 |||| ||||||||||| ||||||||
Sbjct   1888346  GGGATAGACGTTAAAACCGGAAACC   1888370
```

Alignment-based SNP discovery: GATK (1)

- The Genome Analysis Toolkit (GATK) is a software package developed at the Broad Institute to primarily focus on variant discovery and genotyping.
- Input data: BAM files and reference genome
- Required tools: Picard and Samtools
- Code example:

```
java –jar GenomeAnalysisTK.jar \
   -T UnifiedGenotyper \
   -R your_reference \
   -I your_bam \
   -glm BOTH
### BOTH = SNP + INDEL
```

Version 3.7

GATK (2)

VCF (Variant Call Format) output

https://samtools.github.io/hts-specs/VCFv4.2.pdf

isolate 1 ↓

isolate 2 ↓

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	DH10B	MG1655
ref1	89089	.	C	A	782.76	.	...	GT:AD:DP:GQ:MLPSAC:MLPSAF:PL	1:0,18:18:99:1:1.00:781,0	0:27,0:27:99:0:0.00:0,1149
ref1	89103	.	G	C	690.76	.	...	GT:AD:DP:GQ:MLPSAC:MLPSAF:PL	1:0,16:16:99:1:1.00:689,0	0:29,0:29:99:0:0.00:0,1253
ref1	89143	.	A	G	448.76	.	...	GT:AD:DP:GQ:MLPSAC:MLPSAF:PL	1:0,11:11:99:1:1.00:447,0	0:27,0:27:99:0:0.00:0,1165
ref1	89145	.	G	T	405.76	.	...	GT:AD:DP:GQ:MLPSAC:MLPSAF:PL	1:0,10:10:99:1:1.00:404,0	0:28,0:28:99:0:0.00:0,1215

```
GT: AD   : DP: GQ: MLPSAC: MLPSAF: PL
1 : 0,18: 18: 99: 1     : 1.00  : 781,0
```

GT=Genotype (0 or 1)
AD=Allelic depths for the ref and alt alleles
DP=Approximate read depth
GQ=Genotype Quality
MLPSAC=Maximum likelihood expectation (MLE) for the alternate allele count
MLPSAF=Maximum likelihood expectation (MLE) for the alternate allele fraction
PL=Normalized, Phred-scaled likelihoods for genotypes

Prob(0) = 10^(-781/10) = 7.9e-79   Prob(1) = 10^(-0/10) = 1

GATK (3)

- GATK can be used to filter SNPs.

```
java GenomeAnalysisTK.jar \
   -T SelectVariants \
   -R your_reference \
   --variant your_vcf \
   -select 'DP >= 3.0' \
   --restrictAllelesTo BIALLELIC \
   --selectTypeToInclude SNP
```

- Filter variants based on the experimental purpose and genetic features

Falsely discovered SNPs

Can you think about what could result in falsely discovered SNPs using alignment-based SNP methods?

Alignment-based SNP discovery: alignment issues

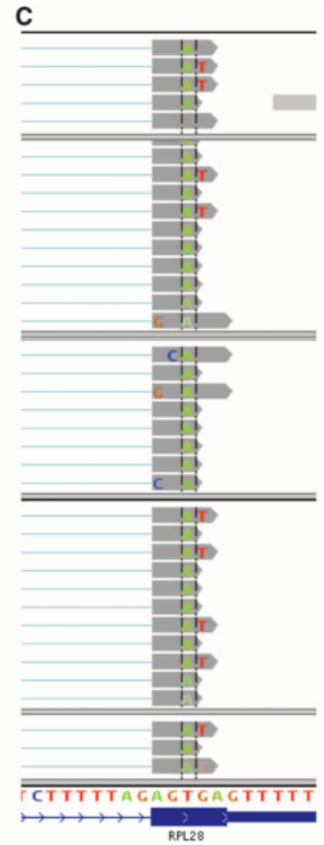
- Misalignments
- Genome duplications
- Highly divergent regions

Examples:

Widespread RNA and DNA Sequence Differences in the Human Transcriptome
Mingyao Li *et al.*
Science **333**, 53 (2011);
DOI: 10.1126/science.1207018

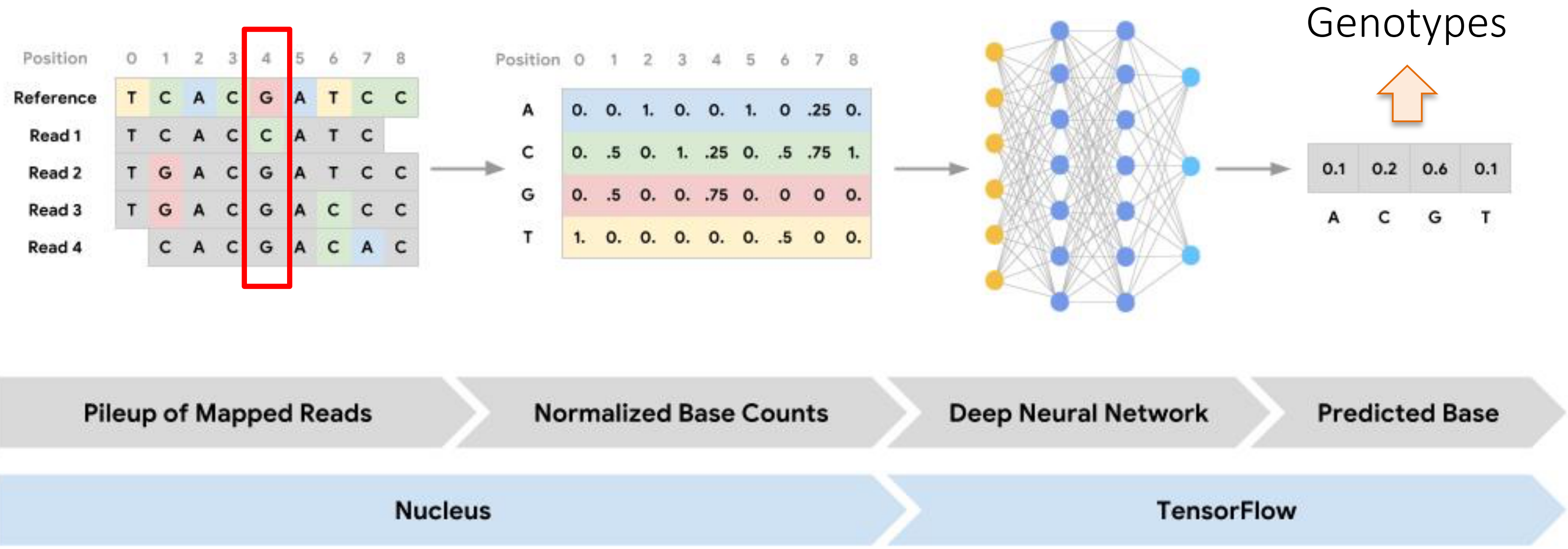
The misalignments of RNA-Seq data or DNA-Seq data led to this discovery

Comment on "Widespread RNA and DNA Sequence Differences in the Human Transcriptome"
Claudia L. Kleinman and Jacek Majewski
Science **335**, 1302 (2012);
DOI: 10.1126/science.1209658



DeepVariant
(alignment-based but with deep learning to infer genotypes)

A universal SNP and small-indel variant caller using deep neural networks



Position	0	1	2	3	4	5	6	7	8
Reference	T	C	A	C	G	A	T	C	C
Read 1	T	C	A	C	C	A	T	C	
Read 2	T	G	A	C	G	A	T	C	C
Read 3	T	G	A	C	G	A	C	C	C
Read 4		C	A	C	G	A	C	A	C

Position	0	1	2	3	4	5	6	7	8
A	0.	0.	1.	0.	0.	1.	0	.25	0.
C	0.	.5	0.	1.	.25	0.	.5	.75	1.
G	0.	.5	0.	0.	.75	0.	0	0	0.
T	1.	0.	0.	0.	0.	0.	.5	0	0.

0.1	0.2	0.6	0.1
A	C	G	T

Figure 1: We formulate consensus-based DNA sequencing error correction as a multiclass classification problem. Using Nucleus, we construct a matrix of normalized base counts in a genomic range. TensorFlow allows us to train a neural network that can predict the correct base at the middle position of the window.

Nature Biotechnology 36:983–987(2018)

Assembly-based SNP discovery

- Cortex (Iqbal *et al*., 2012 Nature Genetics)

de novo assembly and graphic comparison for variant discovery

- Fermi (Li H, 2012 Bioinformatics)

de novo assembly to unitigs* and then alignment to the reference genome for variant discovery

(Conceptually, unitigs are confident contigs)

- Discovar (Neil *et al*., 2014 Nature Genetics)

Region *de novo* assembly to contigs and then alignment to the reference genome for variant discovery

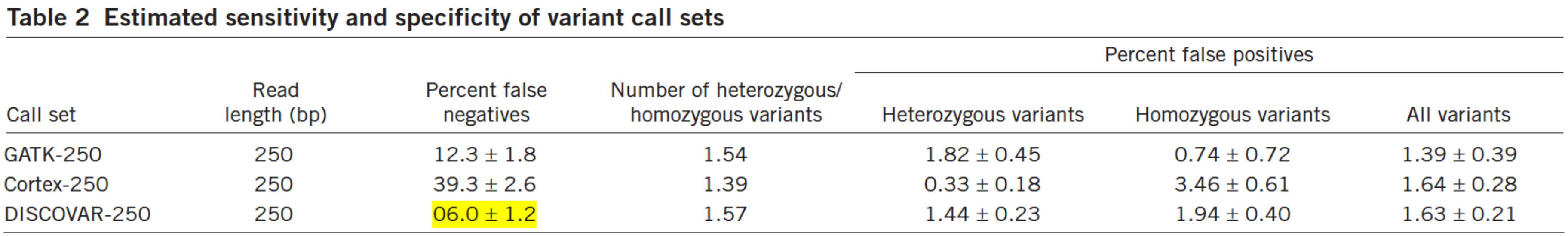


Table 2 Estimated sensitivity and specificity of variant call sets

				Percent false positives		
Call set	Read length (bp)	Percent false negatives	Number of heterozygous/ homozygous variants	Heterozygous variants	Homozygous variants	All variants
GATK-250	250	12.3 ± 1.8	1.54	1.82 ± 0.45	0.74 ± 0.72	1.39 ± 0.39
Cortex-250	250	39.3 ± 2.6	1.39	0.33 ± 0.18	3.46 ± 0.61	1.64 ± 0.28
DISCOVAR-250	250	06.0 ± 1.2	1.57	1.44 ± 0.23	1.94 ± 0.40	1.63 ± 0.21

Variant annotation

Gene coding regions

- ***Synonymous***: changes that do not alter the encoded amino acid
- ***Non-synonymous***

1. ***Missense***: changes that alter encoded amino acid

2. ***Nonsense***: changes that produce a stop codon from an amino acid codon, resulting in a shortened protein

- ***Frameshift*** (caused by insertion/deletion)

Splicing sites

Of an intron, a donor site (5' end of the intron) and an acceptor site (3' end of the intron) are required for splicing.

Variant annotation - SnpEff

SnpEff is a variant annotation and effect prediction tool. It annotates and predicts the effects of variants on genes.

Input data:

- Genome annotation database
- Variant data: VCF file

Running:

```
java –jar snpEff.jar GRCh37.75 my.vcf
```

Cingolani P, et al., DM. Fly (Austin). 2012 Apr-Jun;6(2):80-92.

Detailed effect list from SnpEff

Effect	Note
INTERGENIC	**The variant is in an intergenic region**
UPSTREAM	Upstream of a gene (default length: 5K bases)
UTR_5_PRIME	Variant hits 5'UTR region
UTR_5_DELETED	The variant deletes an exon which is in the 5'UTR of the transcript
START_GAINED	A variant in 5'UTR region produces a three base sequence that can be a START codon.
SPLICE_SITE_ACCEPTOR	The variant hits a splice acceptor site (defined as two bases before exon start, except for the first exon).
SPLICE_SITE_DONOR	The variant hits a Splice donor site (defined as two bases after coding exon end, except for the last exon).
START_LOST	Variant causes start codon to be mutated into a non-start codon.
SYNONYMOUS_START	Variant causes start codon to be mutated into another start codon.
CDS	The variant hits a CDS.
GENE	**The variant hits a gene.**
TRANSCRIPT	The variant hits a transcript.
EXON	The vairant hist an exon.
EXON_DELETED	A deletion removes the whole exon.
NON_SYNONYMOUS_CODING	Variant causes a codon that produces a different amino acid
SYNONYMOUS_CODING	Variant causes a codon that produces the same amino acid
FRAME_SHIFT	**Insertion or deletion causes a frame shift**
CODON_CHANGE	One or many codons are changed
CODON_INSERTION	One or many codons are inserted
CODON_CHANGE_PLUS_CODON_INSERTION	One codon is changed and one or many codons are inserted
CODON_DELETION	One or many codons are deleted
CODON_CHANGE_PLUS_CODON_DELETION	One codon is changed and one or more codons are deleted
STOP_GAINED	**Variant causes a STOP codon**
SYNONYMOUS_STOP	Variant causes stop codon to be mutated into another stop codon.
STOP_LOST	Variant causes stop codon to be mutated into a non-stop codon
INTRON	Variant hist and intron. Technically, hits no exon in the transcript.
UTR_3_PRIME	Variant hits 3'UTR region
UTR_3_DELETED	The variant deletes an exon which is in the 3'UTR of the transcript
DOWNSTREAM	Downstream of a gene (default length: 5K bases)
INTRON_CONSERVED	The variant is in a highly conserved intronic region
INTERGENIC_CONSERVED	The variant is in a highly conserved intergenic region

Summary

- The strategy to generate data for SNP discovery is depended on experimental purpose, genetic features of the population, timetable, and budget.

- A standard approach for SNP discovery is through mapping reads to the reference sequences, thereby identifying variants between reads and reference. The most popular method is GATK.