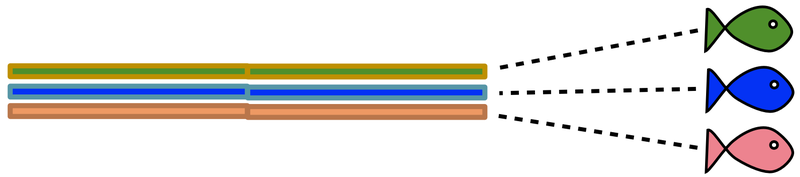
Next-generation DNA sequencing



Calling genotypes for diploid organisms

Regardless of sequencing platform, going from data to genotypes is similar:

1) Determine reference genome to map to, or create a denovo assembly for your data
2) Map reads
3) Call genotypes

Data processing choices have MAJOR downstream influences on your dataset and analyses

4) *Filter, analyze and check, filter again, check again, analyze, filter some more, check some more…*

the large world of…filtering

Pre-variant filtering:

- quality trimming
- removal of low-quality reads

the large world of…filtering

Pre-variant filtering:

- quality trimming
- removal of low-quality reads

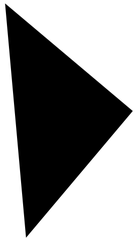
Why???

- Can improve mapping rates
- If the bases are bad, why include them in the first place
- Alternatively: may be filtered out with later due to poor mapping/low confidence genotype calls etc.

the large world of…filtering

Pre-variant filtering:

- quality trimming
- removal of low-quality reads



FASTX toolkit

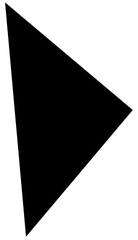
Why???

- Can improve mapping rates
- If the bases are bad, why include them in the first place
- Alternatively: may be filtered out with later due to poor mapping/low confidence genotype calls etc.

the large world of…filtering

Pre-variant filtering:

- quality trimming
- removal of low-quality reads



FASTX toolkit

Uses Phred scores —

Higher Scores (30-60+): Indicate high confidence; these bases are likely correct
Lower Scores (Below 20): Suggest uncertainty; these bases might be errors

Encoded as ASCII characters in FASTQ files

Phred scores

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

(from Wikipedia)

Phred scores

Symbol	Phred Quality Score	Probability of Incorrect Base Call
!	0	1.000
"	1	0.794
#	2	0.631
$	3	0.501
%	4	0.398
&	5	0.316
'	6	0.251
(	7	0.199
)	8	0.158
*	9	0.126
+	10	0.100
,	11	0.079
-	12	0.063
.	13	0.050
/	14	0.040
0	15	0.032
1	16	0.025
2	17	0.020
3	18	0.016
4	19	0.013
5	20	0.010
6	21	0.008
7	22	0.006
8	23	0.005
9	24	0.004
:	25	0.003
;	26	0.002
<	27	0.002
=	28	0.001
>	29	0.001
?	30	0.001
@	31	0.0008
A	32	0.0006
B	33	0.0005
C	34	0.0004
D	35	0.0003
E	36	0.0002
F	37	0.0002
G	38	0.0002
H	39	0.0001
I	40	0.0001

(from Wikipedia)

the large world of…filtering

Post-variant filtering:

Genotype level:

- read depth
- genotype quality
- missing data

Site level:

- missing data
- minor allele frequency (MAF) or minor allele count (MAC)
- Hardy-Weinberg
- Linkage disequilibrium
- total read depth

Variant Call Format (vcf)

Header with metadata

Body with columns/data

```
##fileformat=VCFv4.3
##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">
#CHROM POS      ID         REF    ALT    QUAL  FILTER   INFO                              FORMAT
NA00001         NA00002          NA00003
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   1/1:43:5:.,.
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     0/0:41:3
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     2/2:35:4
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   0/0:61:2
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         1/1:40:3
```

Variant Call Format (vcf)

Body with columns/data

```
#CHROM POS      ID         REF   ALT    QUAL  FILTER   INFO                              FORMAT
NA00001         NA00002          NA00003
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   1/1:43:5:.,.
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     0/0:41:3
```

	Name	Brief description (see the specification for details).
1	CHROM	The name of the sequence (typically a chromosome) on which the variation is being called. This sequence is usually known as 'the reference sequence', i.e. the sequence against which the given sample varies.
2	POS	The 1-based position of the variation on the given sequence.
3	ID	The identifier of the variation, e.g. a dbSNP rs identifier, or if unknown a ".". Multiple identifiers should be separated by semi-colons without white-space.
4	REF	The reference base (or bases in the case of an indel) at the given position on the given reference sequence.
5	ALT	The list of alternative alleles at this position.
6	QUAL	A quality score associated with the inference of the given alleles.
7	FILTER	A flag indicating which of a given set of filters the variation has failed or PASS if all the filters were passed successfully.
8	INFO	An extensible list of key-value pairs (fields) describing the variation. See below for some common fields. Multiple fields are separated by semicolons with optional values in the format: `<key>=<data>[,data]`.
9	FORMAT	An (optional) extensible list of fields for describing the samples. See below for some common fields.
+	SAMPLEs	For each (optional) sample described in the file, values are given for the fields listed in FORMAT

Variant Call Format (vcf)

Body with columns/data

```
#CHROM POS      ID          REF   ALT    QUAL  FILTER   INFO                              FORMAT
NA00001         NA00002           NA00003
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   1/1:43:5:.,.
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     0/0:41:3
```

	Name	Brief description (see the specification for details).
1	CHROM	The name of the sequence (typically a chromosome) on which the variation is being called. This sequen… the sequence against whi…
2	POS	The 1-based position of th…
3	ID	The identifier of the variat… Multiple identifiers should…
4	REF	The reference base (or ba… the given reference seque…
5	ALT	The list of alternative allel…

6	QUAL	A quality score associated with the inference of the given alleles.

Site-level genotype quality score. Phred-scaled (i.e. you can interpret the numbers similarly, but NOT actually Phred scores. This is site-level quality, not individual.

QUAL = 20 → ~1% chance the variant is false
QUAL = 30 → ~0.1% chance
Higher = more confident variant call

Variant Call Format (vcf)

Body with columns/data

```
#CHROM POS     ID        REF    ALT     QUAL  FILTER   INFO                              FORMAT
NA00001        NA00002          NA00003
                                                      P=14;AF=0.5;DB;H2

                                                      P=11;AF=0.017
```

Depends on the variant caller, but can have information about quality and whether quality passes a recommended threshold (e.g. “PASS”)

1	CHROM	The name of the sequence (typically a chromosome) on which the variation is being called. This sequence is usually known as 'the reference sequence', i.e. the sequence against which the given sample varies.
2	POS	The 1-based position of the variation on the given sequence.
3	ID	The identifier of the variation, e.g. a dbSNP rs identifier, or if unknown a ".". Multiple identifiers should be separated by semi-colons without white-space.
4	REF	The reference base (or bases in the case of an indel) at the given position on the given reference sequence.
5	ALT	The list of alternative alleles at this position.

		quality score associated with the inference of the given alleles.
7	FILTER	A flag indicating which of a given set of filters the variation has failed or PASS if all the filters were passed successfully.
8	INFO	An extensible list of key-value pairs (fields) describing the variation. See below for some common fields. Multiple fields are separated by semicolons with optional values in the format: `<key>=<data>[,data]`.
9	FORMAT	An (optional) extensible list of fields for describing the samples. See below for some common fields.
+	SAMPLEs	For each (optional) sample described in the file, values are given for the fields listed in FORMAT

Variant Call Format (vcf)

Body with columns/data

```
#CHROM POS      ID          REF   ALT    QUAL  FILTER   INFO                              FORMAT
NA00001         NA00002             NA00003
20     14370    rs6054257   G     A      29    PASS     NS=3;DP=14;AF=0.5;DB;H2

                                                        P=11;AF=0.017
```

Often includes information about:
1. Site-level read depth (DP) — the sum of individual read depths)
2. Mapping Quality (MQ) — phred-scaled

1	(	the sequence against which the given sample varies.
2	POS	The 1-based position of the variation on the given sequence.
3	ID	The identifier of the variation, e.g. a dbSNP rs identifier, or if unknown a ".". Multiple identifiers should be separated by semi-colons without white-space.
4	REF	The reference base (or bases in the case of an indel) at the given position on the given reference sequence.
5	ALT	The list of alternative alleles at this position.

		quality score associated with the inference of the given alleles.
		flag indicating which of a given set of filters the variation has failed or PASS if ll the filters were passed successfully.
8	INFO	An extensible list of key-value pairs (fields) describing the variation. See below for some common fields. Multiple fields are separated by semicolons with optional values in the format: `<key>=<data>[,data]`.
9	FORMAT	An (optional) extensible list of fields for describing the samples. See below for some common fields.
+	SAMPLEs	For each (optional) sample described in the file, values are given for the fields listed in FORMAT

Variant Call Format (vcf)

Body with columns/data

```
#CHROM POS      ID          REF   ALT    QUAL  FILTER   INFO                              FORMAT
NA00001         NA00002           NA00003
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   1/1:43:5:.,.
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     0/0:41:3
```

FORMAT column includes a Genotype Quality (GQ) measure that is per individual (phred scaled, assigned by the variant caller).

This is an individual genotype level measure!

	Na	
1	CHR(	
2	POS	
3	ID	Multiple identifiers should be separated by semi-colons without white-space.
4	REF	The reference base (or bases in the case of an indel) at the given position on the given reference sequence.
5	ALT	The list of alternative alleles at this position.

		ity score associated with the inference of the given alleles.
		indicating which of a given set of filters the variation has failed or PASS if filters were passed successfully.
		ensible list of key-value pairs (fields) describing the variation. See below me common fields. Multiple fields are separated by semicolons with al values in the format: `<key>=<data>[,data]`.
9	FORMAT	An (optional) extensible list of fields for describing the samples. See below for some common fields.
+	SAMPLEs	For each (optional) sample described in the file, values are given for the fields listed in FORMAT

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT sample1 sample2 sample3

GT:DP:GQ

genotype : read depth : genotype quality

0=ref; 1=alternate (if there is more than one alternate 2, 3)

phred scaled, assigned by the variant caller)

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT sample1 sample2 sample3

GT:DP:GQ

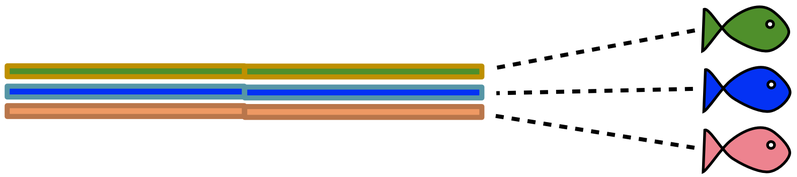
genotype : read depth : genotype quality

0=ref; 1=alternate (if there is more than one alternate 2, 3)

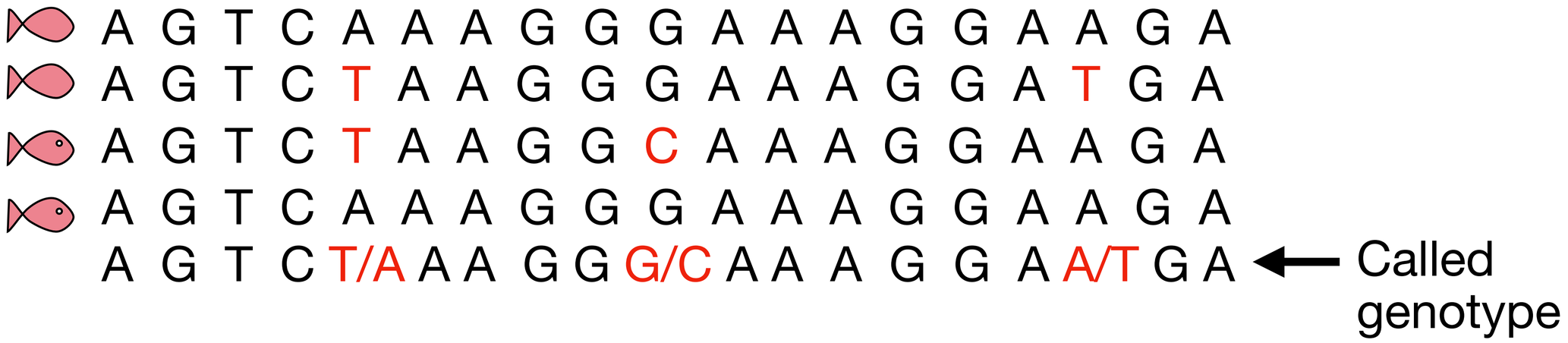
phred scaled, assigned by the variant caller)

chr1 10583 . G A 29.77 PASS DP=14 GT:DP:GQ 0/1:12:35 0/0:14:42 1/1:8:20

Next-generation DNA sequencing



Calling genotypes for diploid organisms



Next-generation DNA sequencing

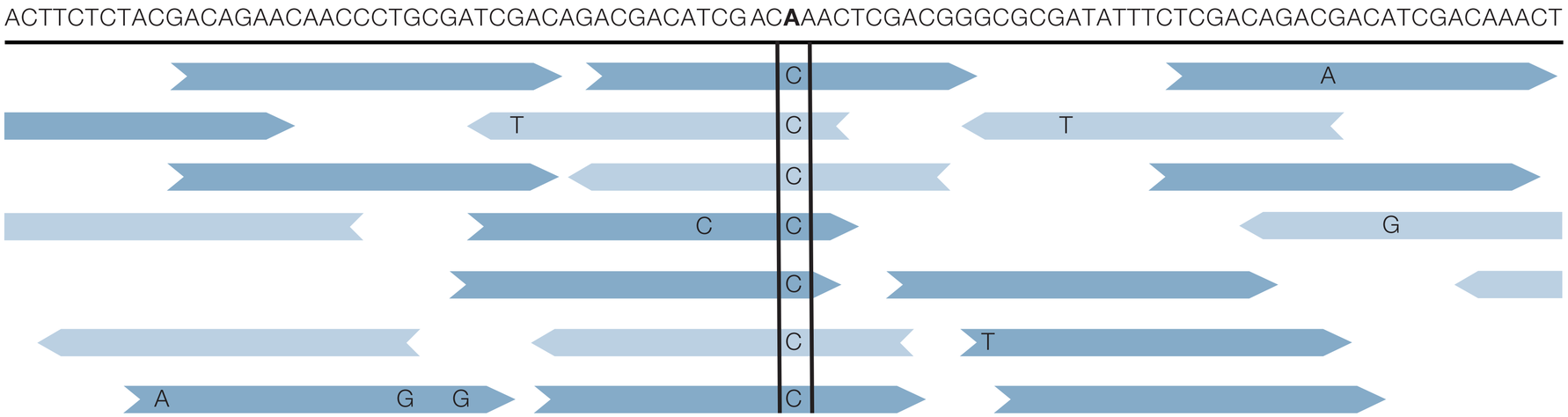


Figure 2.4