Practical Introduction

Sequence Mapping and Assembly

December 8, 2014

Mary Kate Wing University of Michigan Center for Statistical Genetics

Goals of This Session

- Learn basics of sequence data file formats
 FASTQ & BAM
- Raw sequence reads -> aligned sequences
 - Get ready for variant calling
 - Many methods/pipelines, we cover 1
- Evaluate quality of sequence data
- Visualize sequence data to examine reads aligned to particular genomic positions

Session Design

- A few intro slides
 - Introduces you to how to do each of the goals
- Instructions for you to follow
 - Walkthrough of how to produce aligned reads
 - Screenshots with explanations
- Raise your hand if you have any questions/problems
 - Someone will come help

Raw Sequence Reads (FASTQs)

- Standard file format from sequencing
 - Sequencing done as series of reads
 - Not associated with a chromosome/position
 - Reads can be in pairs
 - Typically separate file for 1st/2nd in pair
 - http://en.wikipedia.org/wiki/FASTQ_format

Raw Sequence Reads (FASTQs)

4 lines per read



Raw Sequence Reads (FASTQs)

Base Qualities

- ASCII quality code for each base
 - $33 + \text{phred scale} = 33 + -10\log_{10} \text{e}$
 - e is estimate probability of an incorrect base
 - Lower qualities: special characters/digits
 - ! (Q=0), " (Q=1), # (Q=3), + (Q-10), / (Q=14)
 - 0 (Q=15), 5 (Q=20), 9 (Q=24)
 - Higher qualities (>Q30): alphabetic characters
 - : (Q=25), ? (Q=30), @ (Q=31)
 - A (Q=32), B (Q=33), G (Q=38)
- Will be recalibrated in alignment pipeline
 - By sequencing run/fastq pair
 - Become more accurate

SAM/BAM

- Maps read to Chromosome & Position
 - Spec: <u>http://samtools.github.io/hts-specs/SAMv1.pdf</u>
 - More Info: <u>http://genome.sph.umich.edu/wiki/SAM</u>
- Header lines
 - Each line starts with '@'

Records

- One for each sequence read/FASTQ record
- FASTQ info PLUS Chr/Pos

 Header	@HD	VN:1.3	S0:coord	dinate								
	@SQ	SN:22	LN:51304	4566	AS:NCBI3	37	M5:a718a	caa6135f	dca8357	d5bfe942	11dd	UR
	:file:/h	nome/mkt	rost/seq	shop/got	cloud//	/referenc	e/chr22/	human.g1	.k.v37.cl	hr22.fa		
	@RG	ID:ERR02	13170	SM:HG00	553	LB:g1k-s	c-HG0055	3	PL:ILLU	MINA		
	@RG	ID:ERR0	15764	SM:HG00	553	LB:g1k-s	c-HG0055	3	PL:ILLU	MINA		
	@RG	ID:ERR02	18525	SM:HG00	553	LB:g1k-s	c-HG0055	3-C-6907		PL:ILLU	MINA	
		25.457243		435	22	16300056			39M69H		3646636	
							'%%%\$\$(°9+\$%'%4	<@)\$\$.;5	&@:+\$5(.	AS
	:i:32	NM:i:2	0Q:Z:'%	\$\$\$.\$(,.)	8&>%;*9+;	;;%4<@):	6C;D;@:B	7C(9	RG:Z:ER	R018525	SA:Z:22	2,36
	466074,4	+,60M48S	,0,0;	XS:i:28								
	ERR01317	70.463018	38	97	22	16850138		5	29S50M29	9S	=	36
	809232	19959202	2	AAATGGA	ATCGAATG	GAATTATCO	AATGCAAT	CGAATGGA	ATTATCG	AATGCAAT	CGAATAGA	ATC
	ATCGAATO	GACTCGA	ATGACCCC	FGGGGTAA	GGAGAAGC	CCA	A:=;:9:9	;:1:<;;9):<;<:;:	&91;:9;;	::28;397	/6:;
							0/42)151					
	:Z:ACECO	GHJJGI?K	JHFIKKHI.	JII?LHII.	JLKIJ@LKŀ	IJHLKIIHI	KALKFJIK	KIK?GJIJ	ILKGKJG	=;KKGGBJ	CHA; FBCE	F <f< th=""></f<>
	@JGC=CB6	6B?@B?BC	?B;<;@	RG:Z:ER	R013170	XS:i:36						

(Header	@HD	VN:1.3	S0:co	ordinate								
	Chr info	@SQ	SN:22	LN:51	304566	AS:N	CBI37	M5:a718a	caa6135f	dca8357	d5bfe942	211dd	UR
		:file:/	home/mkt	rost/s	eqshop/go	otcloud,	//referer	nce/chr22/	human.gl	k.v37.c	hr22.fa		
		@RG	ID:ERR0	13170	SM:HG	90553	LB:g1k	-sc-HG0055	3	PL:ILLU	MINA		
		@RG	ID:ERR0	15764	SM:HG	90553	LB:g1k	-sc-HG0055	3	PL:ILLU	MINA		
		@RG	ID:ERR0	18525	SM:HG	90553	LB:g1k	<u>-sc-HG0055</u>	3-C-6907		PL:ILL	JMINA	
		ERR01852	25.45724	33	435	22	1630005	56	0	39M69H	=	3646636	42
							стстстстст			°9+\$%'%4	<@)\$\$.;5	5&@:+\$5(.	AS
		:i:32	NM:i:2	0Q:Z:	'%%%.%(,.8 &>%; '	*9+;,;%4<@)):6C;D;@:B	7C(9	RG:Z:ER	R018525	SA:Z:22	,36
		466074,-	+,60M48S	,0,0;	XS:i:2	28							
		ERR01317	70.46301	88	97	22	1685013	38	5	29S50M2	9S	=	36
		809232	1995920	2	AAATG	GAATCGA	ATGGAATTAT	GAATGCAAT	CGAATGGA	ATTATCG	AATGCAAT	FCGAATAGA	ATC
		ATCGAAT	GGACTCGA	ATGACC	CCTGGGGT	AAGGAGA	AGCCCA	A:=;:9:9	;:1:<;;9	:<;<:;:	§91;:9; ;	;::28;397	6:;
							2+05355//4)						
		:Z:ACECO	GHJJGI?K.	JHFIKK	HIJII?LH	IIJLKIJ(@LKHJHLKIIH	IIKALKFJIK	KIK?GJIJ	ILKGKJG	=;KKGGB3	JCHA; FBCE	F <f< th=""></f<>
		@JGC=CB6	6B?@B?BC	?B;<;@	RG:Z:I	ERR0131	70 XS:i:36	5					

$\left(\right)$	Header	@HD	VN:1.3	S0:coo	rdinate							
	Chrinfo	@SQ	SN:22	LN:513	94566	AS:NCB	I37	M5:a718ac	aa6135fdca835	7d5bfe94	211dd	UR
	Chr info	:file:/	home/mkt	rost/se	qshop/go	tcloud/.	./referenc	e/chr22/h	uman.glk.v37.	chr22.fa	1	
	Read	@RG	ID:ERR0	13170	SM:HG0	0553	LB:g1k-s	c-HG00553	PL:ILL	.UMINA		
		@RG	ID:ERR0	15764	SM:HG0	0553	LB:g1k-s	c-HG00553	PL:ILL	UMINA		
	Group	@RG	ID:ERR0	18525	SM:HG0	0553	LB:g1k-s	c-HG00553	-C-6907	PL:ILL	.UMINA	
		ERR0185	25.45724	33	435	22	16300056	0	39M69H	=	3646636	54 2
		0166378	CACTCTC	TCTCGCT	СТСТСАСТ	стстстст	стстстстс	'%%%\$\$(,	.\$&&%(*9+\$%'%	64 <@)\$\$.;	5&@:+\$5(.	. AS
		:i:32	NM:i:2	0Q:Z:'	****.*(,	.86>%;*9	+;,;%4<@):	6C;D;@:B7	C(9 RG:Z:E	RR018525	5 SA:Z:22	2,36
		466074,	+,60M48S	,0,0;	XS:i:2	8						
		ERR0131	70.46301	88	97	22	16850138	5	29S50M	295	=	36
		809232	1995920	2	AAATGG	AATCGAAT	GGAATTATCG	AATGCAATC	GAATGGAATTATO	GAATGCAA	TCGAATAGA	AATC
		ATCGAAT	GGACTCGA	ATGACCC	CTGGGGTA	AGGAGAAG	CCCA	A:=;:9:9;	:1:<;;9:<;<:;	:&91;:9;	;::28;397	76:;
		;3:6.49	.8/0487,	-686107	04223(/5	33132+	05355//4)5	0/42)1513	16665665/	AS:i:4	0 NM:i:2	0Q
		:Z:ACEC	GHJJGI?K	JHFIKKH	IJII?LHI	IJLKIJ@L	KHJHLKIIHI	KALKFJIKK	IK?GJIJILKGKJ	G=;KKGGE	BJCHA; FBCE	F <f< td=""></f<>
			6B?@B?BC			RR013170					-	
		1 -										

		_											
(Header		@HD	VN:1.3	S0:coo	dinate							
	Chr info	Chr info		SN:22	LN:5130)4566	AS:NCB	[37	M5:a718aca	aa6135fdca83	57d5bfe94	4211dd	UR
	Chr info		:file:/	home/mkt	rost/sec	shop/go	tcloud/.	/referenc	e/chr22/hi	uman.g1k.v37	.chr22.fa	a	
	Read	Í	@RG	ID:ERR0	13170	SM:HG0	0553	LB:g1k-s	c-HG00553	PL:IL	LUMINA		
	Group	┥	@RG	ID:ERR0	15764	SM:HG0	0553	LB:g1k-s	c-HG00553	PL:IL	LUMINA		
C	Group		@RG	ID:ERR0	18525	SM:HG0	0553	LB:g1k-s	<u>c-HG00553</u>	-C-6907	PL:IL	LUMINA	
			ERR0185	25.45724	33	435	22	16300056	0	39M69	H =	364663	64 2
	Record 1		0166378							.\$&&%(*9+\$%'	%4<@)\$\$.	;5&@:+\$5(. AS
			:i:32	NM:i:2	0Q:Z:'9	sss.s(,	.86>%;*9-	+;,;%4<@):	6C;D;@:B70	C(9 RG:Z:	ERR01852	5 SA:Z:22	2,36
			466074,	+,60M48S	,0,0;	XS:i:2	8						
			ERR0131	70.46301	88	97	22	16850138	5	29550	M29S	=	36
			809232	1995920	2	AAATGG	AATCGAAT	GAATTATCG	AATGCAATCO	GAATGGAATTAT	CGAATGCA	ATCGAATAG	AATC
	Record 2		ATCGAAT	GGACTCGA	ATGACCC	TGGGGTA	AGGAGAAG	CCCA	A:=;:9:9;	:1:<;;9:<;<:	;:&91;:9	;;::28;39	76:;
			;3:6.49	.8/0487,	-6861070	4223(/5	33132+0	5355//4)5	0/42)1513	16665665/	AS:i:4	40 NM:i:2	00
										IK?GJIJILKGK			
				6B?@B?BC			RR013170	XS:i:36			-		
			-	-									

Header	@HD VN:1.3 SO:coordinate
Chr info 🔸	<pre>@SQ SN:22 LN:51304566 AS:NCBI37 M5:a718acaa6135fdca8357d5bfe94211dd UR</pre>
	:file:/home/mktrost/seqshop/gotcloud//reference/chr22/human.glk.v37.chr22.fa
Read	<pre>@RG ID:ERR013170 SM:HG00553 LB:g1k-sc-HG00553 PL:ILLUMINA</pre>
Group	<pre>@RG ID:ERR015764 SM:HG00553 LB:g1k-sc-HG00553 PL:ILLUMINA</pre>
Group	<pre>@RG ID:ERR018525 SM:HG00553 LB:g1k-sc-HG00553-C-6907 PL:ILLUMINA</pre>
Read Name	ERR018525.4572433 435 22 16300056 0 39M69H = 36466364 2
from FASTQ	0166378 CACTCTCTCTCGCTCTCTCACTCTCTCTCTCTCTCTCTC '%%%%\$%(,.\$&&%(*9+\$%'%4<@)\$\$.;5&@:+\$5(. AS
	:i:32 NM:i:2 0Q:Z:'%%%%.%(,.8&>%;*9+;,;%4<@):6C;D;@:B7C(9 RG:Z:ERR018525 SA:Z:22,36
(no '@','/1','/2')	466074,+,60M48S,0,0; XS:i:28
	ERR013170.4630188 97 22 16850138 5 29S50M29S = 36
	809232 19959202 AAATGGAATCGAATGGAATTATCGAATGCAATCGAATGGAATTATCGAATGCAATCGAATAGAATC
Record 2	ATCGAATGGACTCGAATGACCCCTGGGGTAAGGAGAAGCCCA A:=;:9:9;:1:<;;9:<;<:;:&91;:9;;::28;3976:;
	;3:6.49.8/0487,-68610704223(/533132+05355//4)50/42)151316665665/ AS:i:40 NM:i:2 00
	:Z:ACECGHJJGI?KJHFIKKHIJII?LHIIJLKIJ@LKHJHLKIIHIKALKFJIKKIK?GJIJILKGKJG=;KKGGBJCHA;FBCEF <f< td=""></f<>
	@JGC=CB6B?@B?BC?B;<;@ RG:Z:ERR013170 XS:i:36

Header	@HD VN:1.3 SO:coordinate
Chr info 🔸	<pre>@SQ SN:22 LN:51304566 AS:NCBI37 M5:a718acaa6135fdca8357d5bfe94211dd UR</pre>
	:file:/home/mktrost/seqshop/gotcloud//reference/chr22/human.glk.v37.chr22.fa
Read	<pre>@RG ID:ERR013170 SM:HG00553 LB:g1k-sc-HG00553 PL:ILLUMINA</pre>
Group	@RG ID:ERR015764 SM:HG00553 LB:g1k-sc-HG00553 PL:ILLUMINA
Group	<pre>@RG ID:ERR018525 SM:HG00553 LB:g1k-sc-HG00553-C-6907 PL:ILLUMINA</pre>
Read Name	ERR018525.4572433 435 22 16300056 0 39M69H = 36466364 2
from FASTQ	0166378 CACTCTCTCTCGCTCTCTCACTCTCTCTCTCTCTCTCTC '%%%%\$%(,.\$&&%(*9+\$%'%4<@)\$\$.;5&@:+\$5(. AS
	:i:32 NM:i:2 0Q:Z:'%%%%.%(, Chromosome/position ;D;@:B7C(9 RG:Z:ERR018525 SA:Z:22,36
(no '@','/1','/2')	466074,+,60M485,0,0; XS:i:28
	ERR013170.4630188 97 22 16850138 5 29S50M29S = 36
	809232 19959202 AAATGGAATCGAATGGAATTGGAATGCAATCGAATGGAATTATCGAATGCAATCGAATGGAATCGAATAGAATC
Record 2	ATCGAATGGACTCGAATGACCCCTGGGGTAAGGAGAAGCCCA A:=;:9:9;:1:<;;9:<;<:;:&91;:9;;::28;3976:;
	;3:6.49.8/0487,-68610704223(/533132+05355//4)50/42)151316665665/ AS:i:40 NM:i:2 00
	:Z:ACECGHJJGI?KJHFIKKHIJII?LHIIJLKIJ@LKHJHLKIIHIKALKFJIKKIK?GJIJILKGKJG=;KKGGBJCHA;FBCEF <f< td=""></f<>
	@JGC=CB6B?@B?BC?B;<;@ RG:Z:ERR013170 XS:i:36

Header	@HD VN:1.3 SO:coordinate
Chr info 🔸	<pre>@SQ SN:22 LN:51304566 AS:NCBI37 M5:a718acaa6135fdca8357d5bfe94211dd UR</pre>
	:file:/home/mktrost/seqshop/gotcloud//reference/chr22/human.g1k Mapping to reference info
Read	<pre>@RG ID:ERR013170 SM:HG00553 LB:g1k-sc-HG00553 P M: match/mismatch</pre>
Group	@RG ID:ERR015764 SM:HG00553 LB:g1k-sc-HG00553 P I: insertion, D: deletion
Gloup	<pre>@RG ID:ERR018525 SM:HG00553 LB:g1k-sc-HG00553-C-6907 PL:ILLUMINA</pre>
	ERR018525.4572433 435 22 16300056 0 39M69H = 36466364 2
from FASTQ	0166378 CACTCTCTCTCGCTCTCTCACTCTCTCTCTCTCTCTC '%%%%\$%(,.\$&&%(*9+\$%'%4<@)\$\$.;5&@:+\$5(. AS
	:i:32 NM:i:2 00:Z:'%%%%.%(, Chromosome/position ;D;@:B7C(9 RG:Z:ERR018525 SA:Z:22,36
(no '@','/1','/2')	466074,+,60M48S,0,0; XS:i:28
	ERR013170.4630188 97 22 16850138 5 29S50M29S = 36
	809232 19959202 AAATGGAATCGAATGGAATTATCGAATGCAATCGAATGGAATTATCGAATGCAATCGAATAGAATC
Record 2	ATCGAATGGACTCGAATGACCCCTGGGGTAAGGAGAAGCCCA A:=;:9:9;:1:<;;9:<;<:;:&91;:9;;::28;3976:;
	;3:6.49.8/0487,-68610704223(/533132+05355//4)50/42)151316665665/ AS:i:40 NM:i:2 0Q
	:Z:ACECGHJJGI?KJHFIKKHIJII?LHIIJLKIJ@LKHJHLKIIHIKALKFJIKKIK?GJIJILKGKJG=;KKGGBJCHA;FBCEF <f< td=""></f<>
	@JGC=CB6B?@B?BC?B;<;@ RG:Z:ERR013170 XS:i:36

Header	@HD VN:1.3 SO:coordinate
Chr info 🗲	<pre>@SQ SN:22 LN:51304566 AS:NCBI37 M5:a718acaa6135fdca8357d5bfe94211dd UR</pre>
	:file:/home/mktrost/seqshop/gotcloud//reference/chr22/human.glk Mapping to reference info
Read	<pre>@RG ID:ERR013170 SM:HG00553 LB:g1k-sc-HG00553 P M: match/mismatch</pre>
Group	<pre>@RG ID:ERR015764 SM:HG00553 LB:g1k-sc-HG00553 P I: insertion, D: deletion</pre>
Group	<pre>@RG ID:ERR018525 SM:HG00553 LB:g1k-sc-HG00553-C-6907 PL:ILLUMINA</pre>
Read Name	ERR018525.4572433 435 22 16300056 0 39M69H = <u>36466364</u> 2
from FASTQ	0166378 CACTCTCTCTCGCTCTCTCACTCTCTCTCTCTCTCTC '%%%%\$%(,.\$&&(*9+\$%'%4<@)\$\$.;5&@:+\$5(. AS
	:i:32 NM:i:2 00:Z:'%%%.%(, Chromosome/position ;D;@:B7C(9 RG: paired-end, mate chr/pos
(no '@','/1','/2')	466074,+,60M48S,0,0; XS:i:28
	ERR013170.4630188 97 22 16850138 5 29S50M29S = 36
	809232 19959202 AAATGGAATCGAATGGAATTATCGAATGCAATCGAATGGAATTATCGAATGCAATCGAATAGAATC
Record 2	ATCGAATGGACTCGAATGACCCCTGGGGTAAGGAGAAGCCCA A:=;:9:9;:1:<;;9:<;<:;:&91;:9;;::28;3976:;
	;3:6.49.8/0487,-68610704223(/533132+05355//4)50/42)151316665665/ AS:i:40 NM:i:2 00
	:Z:ACECGHJJGI?KJHFIKKHIJII?LHIIJLKIJ@LKHJHLKIIHIKALKFJIKKIK?GJIJILKGKJG=;KKGGBJCHA;FBCEF <f< td=""></f<>
	@JGC=CB6B?@B?BC?B;<;@ RG:Z:ERR013170 XS:i:36
	630C-CDOD:6D:DC:D, 7,6 NO.2. LINOIDITO X3.1.30

										-
Header	@HD	VN:1.3	S0:coo	rdinate						
Chrinfo	@SQ	SN:22	LN:513	04566	AS:NCB	I37 M5:	a718acaa613	35fdca8357d5b [.]	fe94211dd	UR
Chr info 🗲	:file:/	/home/mkt	rost/se	qshop/go	tcloud/.	./reference/c	hr22/human	.glk Mapping to	o reference info)
Read	@RG	ID:ERR0	13170	SM:HG0	0553	LB:g1k-sc-H	G00553	P M: mat	ch/mismatch	
	@RG	ID:ERR0	15764	SM:HG0	0553	LB:g1k-sc-H	G00553	P I: insertio	on, D: deletion	
Group	ORG	ID:ERR0	18525	SM:HG0	0553	LB:g1k-sc-H	G00553-C-69	907PL	: ILLUMINA	
Read Name	ERR0185	525.45724	33	435	22	16300056	0	39M69H =	3646636	64 2
	0166378	3 CACTCTC	тстсбст	СТСТСАСТ	стстстст	CTETCTCTC '%%	%%\$%(\$&&	s(*9+\$%'%4<@)	\$\$.; 5& @:+\$5(.	. AS
from FASTQ	:i:32								end, mate chr/p	
(no'@'''/1'''/2')			0Q:Z:'		Chromos	ome/position ;				
(no'@'''/1'''/2')	466074,	NM:i:2	0Q:Z:' ,0,0;	\$ ** *.*(,	Chromos					
(no '@','/1','/2')	466074,	NM:i:2 . <u>+,60M48S</u> 170.46301	0Q:Z:' ,0,0; 88	%%%.%(, XS:i:2 97	Chromos 8 22	ome/position ; 16850138	D;@:B7C(9 5	RG: paired-	end, mate chr/p =	oos 36
(no '@','/1','/2') Sequence	466074, ERR0131 809232	NM:i:2 <u>+,60M485</u> 170.46301 1995920	0Q:Z:' ,0,0; 88 2	%%%.%(, XS:i:2 97 AAATGG	Chromos 8 22 AATCGAAT	come/position ; 16850138 GGAATTATCGAAT	D;@:B7C(9 5 GCAATCGAATC	RG: paired-0	end, mate chr/p = GCAATCGAATAGA	oos 36 AATC
(no '@','/1','/2') Sequence	466074, ERR0131 809232 ATCGAAT	NM:i:2 +,60M48S 170.46301 1995920 TGGACTCGA	0Q:Z:' ,0,0; 88 <u>2</u> ATGACCC	%%%%.%(, XS:i:2 97 AAATGG CTGGGGTA	Chromos 8 22 AATCGAAT AGGAGAAG	come/position ; 16850138 GGAATTATCGAAT	D;@:B7C(9 5 GCAATCGAAT(;:9:9;:1:<	RG: paired- 29S50M29S GGAATTATCGAAT(;;9:<;<:;:&91	end, mate chr/p = GCAATCGAATAGA	36 36 AATC 76:;
(no '@','/1','/2') Sequence	466074, ERR0131 809232 ATCGAAT ;3:6.49	NM:i:2 +,60M48S 170.46301 1995920 GGACTCGA 9.8/0487,	0Q:Z:' ,0,0; 88 2 ATGACCC -686107	%%%%.%(, <u>XS:i:2</u> 97 AAATGG CTGGGGTA 04223(/5	Chromos 8 22 AATCGAAT AGGAGAAG 33132+	GGAATTATCGAAT CCCA A:= 05355//4)50/4	D;@:B7C(9 5 GCAATCGAATC ;:9:9;:1:< 2)15131666	RG: paired- 29S50M29S GGAATTATCGAAT ;;9:<;<:;:&91 5665/ AS	end, mate chr/p = GCAATCGAATAGA ;:9;;::28;397 :i:40 NM:i:2	36 <u>36</u> AATC 76:; 00
(no '@','/1','/2') Sequence (from FASTQ)	466074, ERR0131 809232 ATCGAAT ;3:6.49 :Z:ACEC	NM:i:2 +,60M48S 170.46301 1995920 GGACTCGA 9.8/0487,	0Q:Z:' ,0,0; 88 2 ATGACCC -686107 JHFIKKH	%%%%.%(, <u>XS:i:2</u> 97 AAATGG CTGGGGTA 04223(/5 IJII?LHI	Chromos 8 22 AATCGAAT AGGAGAAG 33132+	GGAATTATCGAAT GGAATTATCGAAT CCCA A:= 05355//4)50/4 KHJHLKIIHIKAL	D;@:B7C(9 5 GCAATCGAATC ;:9:9;:1:< 2)15131666	RG: paired- 29S50M29S GGAATTATCGAAT(;;9:<;<:;:&91	end, mate chr/p = GCAATCGAATAGA ;:9;;::28;397 :i:40 NM:i:2	36 <u>36</u> AATC 76:; 00

											-
Header	@HD	VN:1.3	S0:coo	rdinate							
Chr info 🗲	@SQ	SN:22	LN:513	04566	AS:NCB	I37 M5	:a718acaa6	135fdca835	7d5bfe94	4211dd	UR
	:file:	/home/mkt	rost/se	qshop/go	tcloud/.	./reference/	chr22/huma	n.glk Map	ping to re	ference info	·
Read	@RG	ID:ERR0	13170	SM:HG0	0553	LB:g1k-sc-	HG00553	P N	I: match/n	nismatch	
Group	@RG	ID:ERR0	15764	SM:HG0	0553	LB:g1k-sc-H	HG00553	P I: ir	sertion, I	D: deletion	
Group	ORG	ID:ERR0	18525	SM:HG0	0553	LB:g1k-sc-	HG00553-C-	6907	PL:IL	LUMINA	
Read Name	ERR018	525.45724	33	435	22	16300056	Θ	39M69H	=	3646636	<u>i4</u> 2
from FASTQ	016637	8 CACTCTC	TCTCGCT	CTCTCACT	стстстст	CTCTCTCTC '%	۶۶۶ ۶۶(,. \$&	&%(*9+\$%'%	4<@)\$\$.	;5&@:+\$5(.	AS
	:i:32	NM:i:2	0Q:Z:'			ome/position	;D;@:B7C(9	RG: pa	aired-end	, mate chr/p	os
(no '@','/1','/2')		NM:i:2 ,+,60M48S			Chromos		;D;@:B7C(9	RG: pa	aired-end	, mate chr/p	os
	466074		,0,0;	%%%%.%(, <u>XS:i:2</u> 97	Chromos 8 22	ome/position	5	29S50M	295	=	36
	466074	<u>,+,60M48S</u> 170.46301	,0,0; 88	%%%%.%(, <u>XS:i:2</u> 97	Chromos 8 22	ome/position	5	29S50M	295	=	36
(no '@','/1','/2') Sequence (from FASTQ)	466074 ERR013 809232 ATCGAA	,+,60M48S 170.46301 1995920 TGGACTCGA	,0,0; 88 2 ATGACCC	<pre>%%%.%(, XS:i:2 97 AAATGG CTGGGGTA</pre>	Chromos 8 22 AATCGAATC	GGAATTATCGAAT	5 TGCAATCGAA =;:9:9;:1:	29S50M TGGAATTATC <;;9:<;;	129S GAATGCA	= ATCGAATAGA	36 ATC
(no '@','/1','/2') Sequence (from FASTQ)	466074 ERR013 809232 ATCGAA ;3:6.4	,+,60M48S 170.46301 1995920 TGGACTCGA 9.8/0487,	,0,0; 88 2 ATGACCC -686107	%%%.%(, <u>XS:i:2</u> 97 AAATGG CTGGGGTA 04223(/5	Chromos 22 AATCGAATC AGGAGAAG	CCCA A:= 05355//4)50/4	5 TGCAATCGAA =;:9:9;:1: 42)1513166	29S50M TGGAATTATC <;;9:<;<:; 65665/	29S GAATGCA :&91;:9 AS:i:4	= ATCGAATAGA ;;::28;397 40 NM:1:2	36 ATC 76:; 0Q
(no '@','/1','/2') Sequence (from FASTQ)	466074 ERR013 809232 ATCGAA ;3:6.4	,+,60M48S 170.46301 1995920 TGGACTCGA 9.8/0487,	,0,0; 88 2 ATGACCC -686107	%%%.%(, <u>XS:i:2</u> 97 AAATGG CTGGGGTA 04223(/5	Chromos 22 AATCGAATC AGGAGAAG	GGAATTATCGAAT	5 TGCAATCGAA =;:9:9;:1: 42)1513166	29S50M TGGAATTATC <;;9:<;<:; 65665/	29S GAATGCA :&91;:9 AS:i:4	= ATCGAATAGA ;;::28;397 40 NM:1:2	36 ATC 76:; 0Q
(no '@','/1','/2') Sequence (from FASTQ) Recalibrated	466074 ERR013 809232 ATCGAA ;3:6.4 :Z:ACE	,+,60M48S 170.46301 1995920 TGGACTCGA 9.8/0487,	,0,0; 88 2 ATGACCC -686107 JHFIKKH	%%%.%(, XS:i:2 97 AAATGG CTGGGGTA 04223(/5 IJJII?LHI	Chromos 22 AATCGAATC AGGAGAAG 33132+0 IJLKIJ@L	CCCA A:= 05355//4)50/4	5 TGCAATCGAA =;:9:9;:1: 42)1513166	29S50M TGGAATTATC <;;9:<;<:; 65665/	29S GAATGCA :&91;:9 AS:i:4	= ATCGAATAGA ;;::28;397 40 NM:1:2	36 ATC 76:; 00

Viewing SAM/BAM Files

• Samtools - what we will use today

- <u>http://samtools.sourceforge.net/</u>
- view

read group, library, MAPQ >, region

• tview

text alignment viewer - visualize reads by position

- BamUtil (Michigan tool)
 - <u>http://genome.sph.umich.edu/wiki/BamUtil</u>
 - Lot's of SAM/BAM tools
- Other tools
 - Picard, GATK, BamTools

High Quality BAMs from FASTQs

- 1. Map FASTQ reads to reference genome
 - Identify most likely chromosome/position for each read
- 2. Remove Duplicates
 - Remove duplicates so they are not counted multiple times as evidence for variant
- 3. Recalibrate Base Qualities
 - Improve qualities from sequencer

High Quality BAMs from FASTQs

Many tools & best practices to choose from
Our solution:

Genomes on the Cloud (GotCloud)

- Sequence analysis pipelines
 - You don't need to know the details of individual components
 - Automates steps for you

What is Genomes On The Cloud?

- Integrative
- Seamless
- Robust
- Scalable
- Alignment, QC, Variant Calling, Phasing Requires only simple configuration files ..against unexpected failures & stops ..to many thousands of genomes

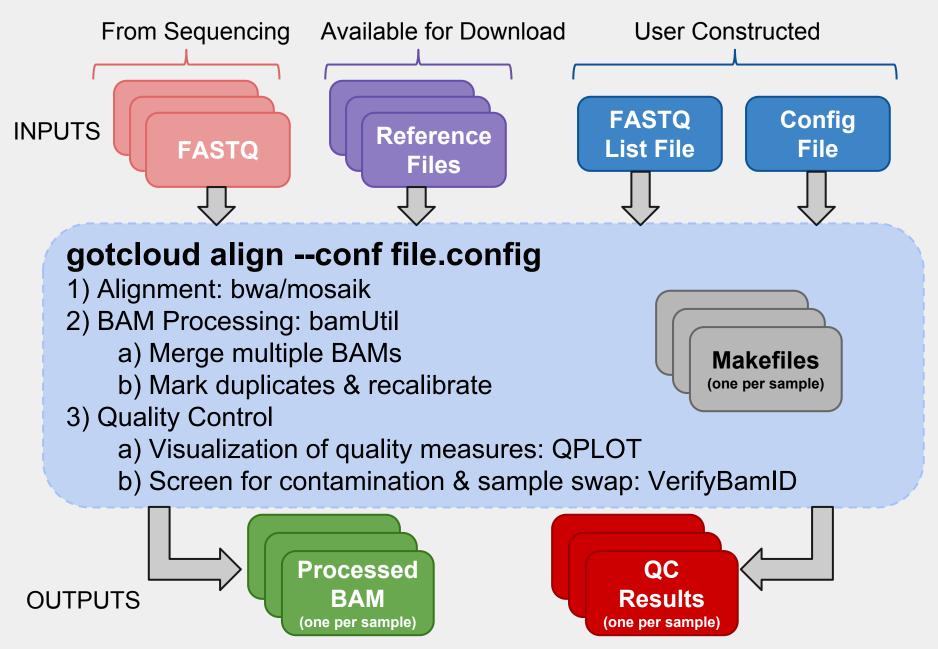
- GotCloud also provides
 - Set of many useful software tools
 - Software library (C++) for sequence analysis

GotCloud Pipelines

Robust parallelization

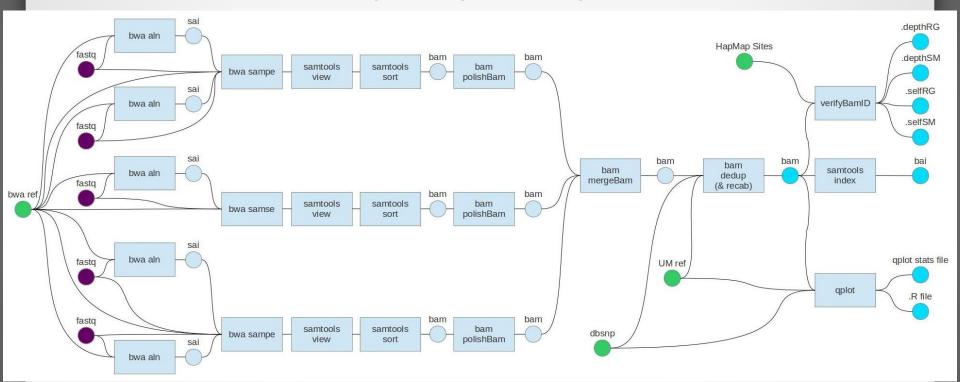
- Automatically partitions
 - Alignment: by sample
 - Others: by region
- Takes advantage of clusters
 - Supports MOSIX, SLURM, SGE, PBS
 - Can setup a cluster on Amazon
 - Amazon Machine Image (AMI) available
- via GNU make
 - Reliable and fault-tolerant
 - Restart where it stopped

GotCloud Alignment Pipeline Overview



What are all of the steps in GotCloud align?

For each sample (using BWA):



Aren't you glad you don't have to worry about and manage each step?

GotCloud needs to know about each FASTQ

- Where to find it
- Sample name
 - Each sample can have multiple FASTQs
 - 1 FASTQ only has a single sample

• Format

- Tab delimited
- Header line
- One line per single-end
- One line per paired-end

* Spacing adjusted for easier reading (just a single tab between columns)

Header Row

	↑	
SAMPLE FASTQ1		FASTQ2
HG00551 fastq/H	G00551.SRR190851.fastq	
HG00551 fastq/H	G00551.SRR190851_1.fastq	<pre>fastq/HG00551.SRR190851_2.fastq</pre>
HG00553 fastq/H	G00553.ERR013170.fastq	
HG00553 fastq/H	G00553.ERR013170_1.fastq	<pre>fastq/HG00553.ERR013170_2.fastq</pre>
HG00553 fastq/H	G00553.ERR015764.fastq	
HG00553 fastq/H	G00553.ERR015764_1.fastq	<pre>fastq/HG00553.ERR015764_2.fastq</pre>
HG00553 fastq/H	G00553.ERR018525.fastq	
HG00553 fastq/H	G00553.ERR018525 1.fastq	<pre>fastq/HG00553.ERR018525 2.fastq</pre>
HG00640 fastq/H	G00640.ERR013174.fastq	
HG00640 fastq/H	G00640.ERR013174 1.fastq	<pre>fastq/HG00640.ERR013174 2.fastq</pre>
HG00640 fastq/H	G00640.ERR015768.fastq	
HG00640 fastq/H	G00640.ERR015768 1.fastq	<pre>fastq/HG00640.ERR015768 2.fastq</pre>
HG00640 fastq/H	G00640.ERR018527.fastq	
HG00640 fastq/H	G00640.ERR018527 1.fastq	<pre>fastq/HG00640.ERR018527 2.fastq</pre>
HG00641 fastq/H	G00641.SRR069531.fastq	
		<pre>fastq/HG00641.SRR069531 2.fastq</pre>
		• – •

* Spacing adjusted for easier reading (just a single tab between columns)

Header Row

SAMPLE FASTQ1 FASTQ2	\Box
HG00551 fastq/HG00551.SRR190851.fastq .	
HG00551 fastq/HG00551.SRR190851_1.fastq fastq/HG00551.SRR190851_2.fastq	
HG00553 fastq/HG00553.ERR013170.fastq .	
HG00553 fastq/HG00553.ERR013170_1.fastq fastq/HG00553.ERR013170_2.fastq	
HG00553 fastq/HG00553.ERR015764.fastq .	
HG00553 fastq/HG00553.ERR015764_1.fastq fastq/HG00553.ERR015764_2.fastq	
HG00553 fastq/HG00553.ERR018525.fastq .	
HG00553 fastq/HG00553.ERR018525_1.fastq fastq/HG00553.ERR018525_2.fastq	
HG00640 fastq/HG00640.ERR013174.fastq .	
HG00640 fastq/HG00640.ERR013174_1.fastq fastq/HG00640.ERR013174_2.fastq	
HG00640 fastq/HG00640.ERR015768.fastq .	
HG00640 fastq/HG00640.ERR015768_1.fastq fastq/HG00640.ERR015768_2.fastq	
HG00640 fastq/HG00640.ERR018527.fastq .	
HG00640 fastq/HG00640.ERR018527_1.fastq fastq/HG00640.ERR018527_2.fastq	
HG00641 fastq/HG00641.SRR069531.fastq .	
HG00641 fastq/HG00641.SRR069531_1.fastq fastq/HG00641.SRR069531_2.fastq	

Groups all FASTQs for a sample in a single BAM

* Spacing adjusted for easier reading (just a single tab between columns)

Header Row

	SAMPLE	FASTQ1	FASTQ2			
	HG00551	fastq/HG00551.SRR190851.fastq				
		<pre>fastq/HG00551.SRR190851_1.fastq</pre>	<pre>fastq/HG00551.SRR190851_2.fastq</pre>			
	HG00553	fastq/HG00553.ERR013170.fastq				
	HG00553	<pre>fastq/HG00553.ERR013170_1.fastq</pre>	<pre>fastq/HG00553.ERR013170_2.fastq</pre>			
	HG00553	fastq/HG00553.ERR015764.fastq				
	HG00553	<pre>fastq/HG00553.ERR015764_1.fastq</pre>	<pre>fastq/HG00553.ERR015764_2.fastq</pre>			
	HG00553	fastq/HG00553.ERR018525.fastq				
	HG00553	<pre>fastq/HG00553.ERR018525_1.fastq</pre>	<pre>fastq/HG00553.ERR018525_2.fastq</pre>			
	HG00640	fastq/HG00640.ERR013174.fastq				
	HG00640	<pre>fastq/HG00640.ERR013174_1.fastq</pre>	fastq/HG00640.ERR013174_2.fastq			
	HG00640	fastq/HG00640.ERR015768.fastq				
		<pre>fastq/HG00640.ERR015768_1.fastq</pre>	fastq/HG00640.ERR015768_2.fastq			
	HG00640	fastq/HG00640.ERR018527.fastq				
	HG00640	<pre>fastq/HG00640.ERR018527_1.fastq</pre>	<pre>fastq/HG00640.ERR018527_2.fastq /</pre>			
	HG00641	fastq/HG00641.SRR069531.fastq				
	HG00641	<pre>fastq/HG00641.SRR069531_1.fastq</pre>	<pre>fastq/HG00641.SRR069531_2.fastq</pre>			
1						

Groups all FASTQs for a sample in a single BAM

Multiple FASTQs for 1 sample

* Spacing adjusted for easier reading (just a single tab between columns)

Header Row

Ť				
ſ	SAMPLE	FASTQ1 FASTQ2		
	HG00551	fastq/HG00551.SRR190851.fastq 🕟		
	HG00551	fastq/HG00551.SRR190851_1.fastq		
	HG00553	fastq/HG00553.ERR013170.fastq :: means single-end		
	HG00553	fastq/HG00553.ERR013170_1.fastq filename means 2nd in pair 2.fastq		
	HG00553	fastq/HG00553.ERR015764.fastq		
	HG00553	fastq/HG00553.ERR015764_1.fastq fastq/HG00553.ERR015764_2.fastq		
	HG00553	fastq/HG00553.ERR018525.fastq .		
	HG00553	fastq/HG00553.ERR018525_1.fastq fastq/HG00553.ERR018525_2.fastq		
1	HG00640	fastq/HG00640.ERR013174.fastq .		
	HG00640	<pre>fastq/HG00640.ERR013174_1.fastq fastq/HG00640.ERR013174_2.fastq</pre>		
	HG00640	<pre>fastq/HG00640.ERR015768.fastq .</pre>		
	HG00640	<pre>fastq/HG00640.ERR015768_1.fastq fastq/HG00640.ERR015768_2.fastq</pre>		
	HG00640	<pre>fastq/HG00640.ERR018527.fastq .</pre>		
	HG00640	fastq/HG00640.ERR018527_1.fastq fastq/HG00640.ERR018527_2.fastq /		
	HG00641	fastq/HG00641.SRR069531.fastq .		
	HG00641	<pre>fastq/HG00641.SRR069531_1.fastq fastq/HG00641.SRR069531_2.fastq</pre>		

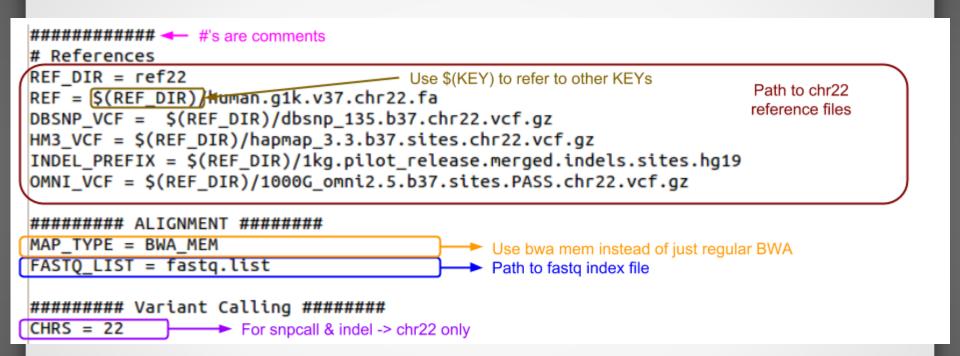
Groups all FASTQs for a sample in a single BAM

Multiple FASTQs for 1 sample

FASTQ List File: Optional RG Fields

Field Name	Description	Default
RGID	Read Group ID - unique for each run	Derived from first line of FASTQ or incrementing numbers
LIBRARY	Separates FASTQs for a sample that were prepared separately	SAMPLE
PLATFORM	Sequencing Platform	ILLUMINA
CENTER	Sequencing Center Useful if multiple centers	unknown
MERGE_NAME	Group FASTQs into a single BAM file Only need if want multiple BAMs for a sample	SAMPLE

User Created Input: Configuration



User Created Input: Configuration

GotCloud Quality Control:

Sample Contamination/Swap (by VerifyBamID)

- Genotype-free estimate of contamination
 - 0-1 scale, the lower, the better
 - 'FREEMIX' column < 0.03
 - http://genome.sph.umich.edu/wiki/VerifyBamID# A_guideline_to_interpret_output_files

FREEMIX FREELK1 FREELK0 FREE_RH FREE_RA CHIPMIX

NA

NA

NA

0.00000 955.99 955.99

Estimate of contamination with genotypes

AVG DP

0.18

- 0-1 scale, the lower, the better
- 'CHIPMIX' column

20056

CHIP ID #SNPS

NA

#SEO ID RG

HG00551 ALL

We don't have this in our tutorial

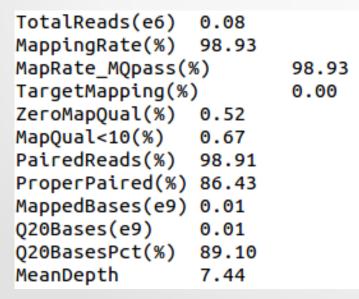
#READS

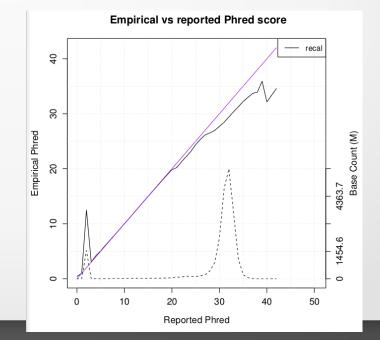
3654

GotCloud Quality Control:

Quality Metrics (by QPLOT)

- .stats file contains metrics, including
 mapping rate, coverage, % high quality bases
- .R file that generates a .pdf of plots
 - Empirical vs reported Phred score





Try it yourself

http://genome.sph.umich.edu/wiki/SeqShop: Sequence Mapping and Assembly Practical

- Interested in GotCloud?
 - <u>http://genome.sph.umich.edu/wiki/GotCloud</u>
 - Join the mailing list:
 - http://groups.google.com/group/GotCloud