SNV and SV calling

Guillaume Bourque

Dept. Human Genetics, McGill University
McGill University and Genome Quebec Innovation Center
Canadian Center for Computational Genomics (C3G)

@guilbourque

Dec 5th 2017

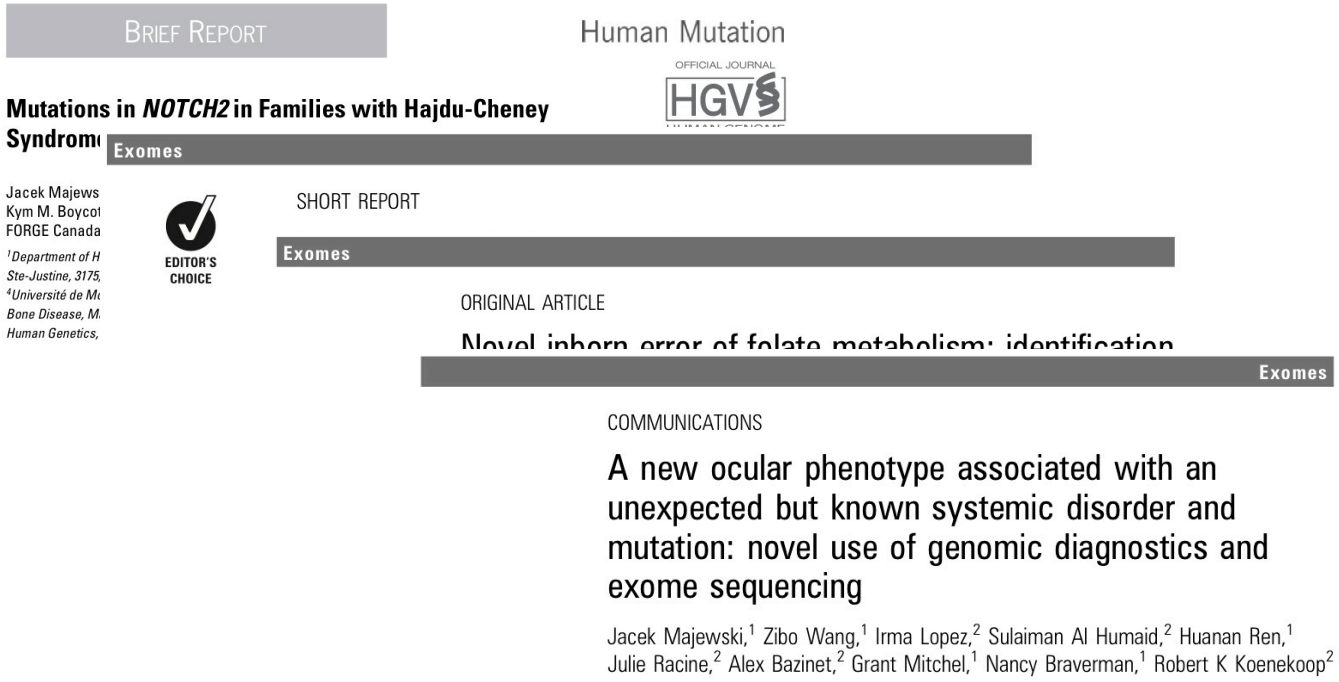
bioinformatics.ca

McGill

Applications of NGS

- ***Denovo* genome sequencing** (Human, model organisms, extreme genomes, everything…)
- **Genome re-sequencing** (Hapmap, 1000 Genome projects, mendelian diseases, cancer, etc.)
- **Functional genomics** (protein-DNA interactions, transcriptome sequencing, chromatin configuration assays …)

Exome sequencing



BRIEF REPORT

Human Mutation

OFFICIAL JOURNAL HGVS

Mutations in *NOTCH2* in Families with Hajdu-Cheney Syndrom

Exomes

Jacek Majews
Kym M. Boycot
FORGE Canada

[1]Department of H
Ste-Justine, 3175,
[4]Université de M
Bone Disease, M.
Human Genetics,

EDITOR'S CHOICE

SHORT REPORT

Exomes

ORIGINAL ARTICLE

Novel inborn error of folate metabolism: identification

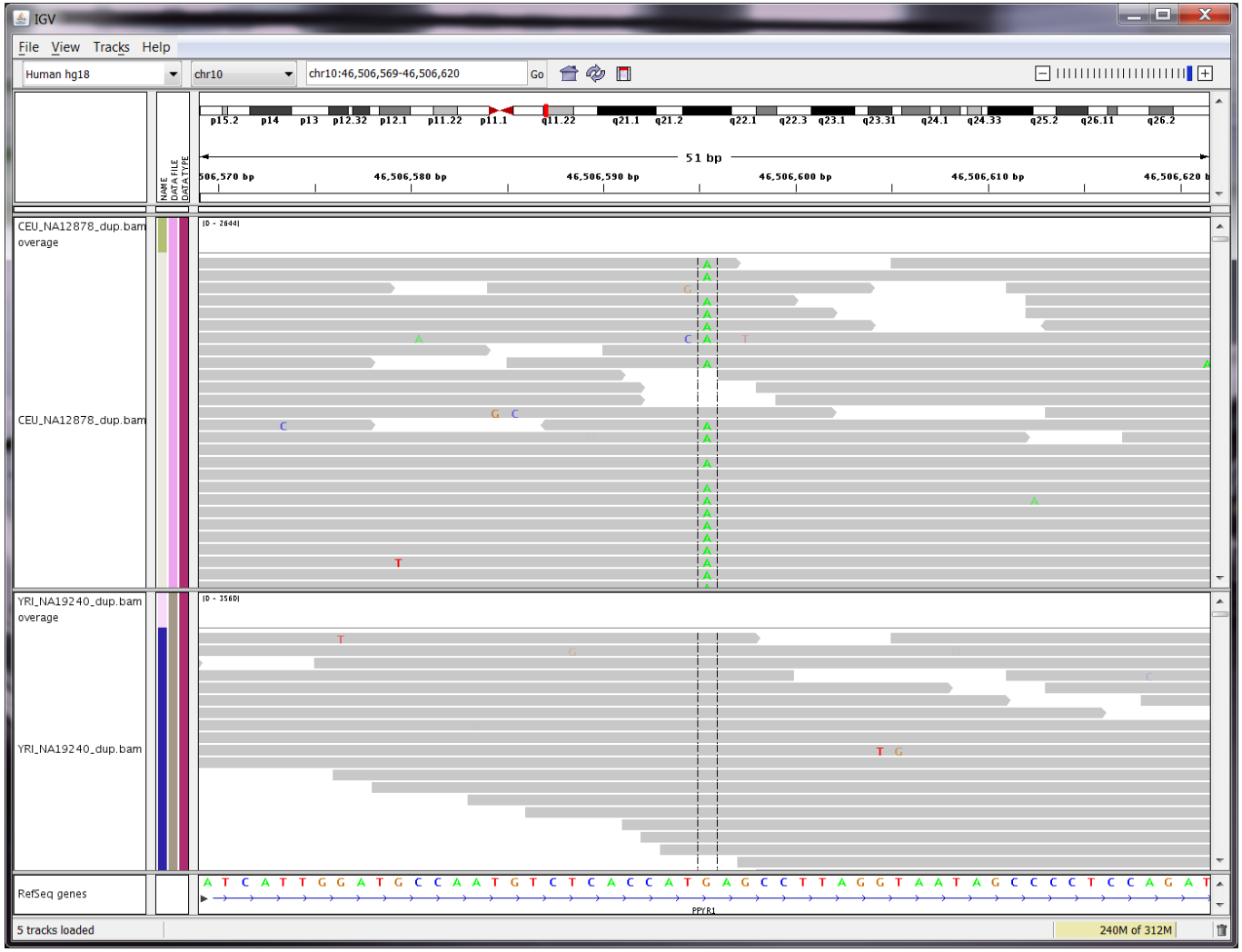
Exomes

COMMUNICATIONS

A new ocular phenotype associated with an unexpected but known systemic disorder and mutation: novel use of genomic diagnostics and exome sequencing

Jacek Majewski,[1] Zibo Wang,[1] Irma Lopez,[2] Sulaiman Al Humaid,[2] Huanan Ren,[1] Julie Racine,[2] Alex Bazinet,[2] Grant Mitchel,[1] Nancy Braverman,[1] Robert K Koenekoop[2]

McGill



Michael Stromberg, bioinformatics.ca

McGill

{

Standard data formats

- Fasta (text format, id + sequence)
- Fastq (text format, id + sequence + qualities)
- SAM (text format, id + seq + qual + alignment)
- BAM (same as SAM but binary encoded)
- VCF (variant call format)
- …

Base quality and Phred scores

- *Q_sanger* = -10 log_10 (*p*)

Where *Q* is the quality and *p* is the probability of the base being incorrect.

Encoding:

- *Phred33* = *Q_sanger* + 33
- *Phred64* = *Q_sanger* + 64 [before Illumina 1.8]

What is a base quality?

Base Quality	P_{error}(obs. base)
3	50 %
5	32 %
10	10 %
20	1 %
30	0.1 %
40	0.01 %

Genetic Variation **Discovery** bioinformatics.ca

From MICHAEL STRÖMBERG

McGill

Fastq example

```
@EAS54_6
CCCTTCTT
+
;;3;;;;;
@EAS54_6
TTGGCAGG
+
;;;;;;;;
```

http://maq.

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

SAM format

The SAM Format Specification (v1.4-r985)

The SAM Format Specification Working Group

September 7, 2011

1 The SAM Format Specification

SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to the alignments. Header lines start with '@', while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information.

samtools.sourceforge.net/SAM1.pdf

SAM example

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1
+r002
+r003
+r004
-r003
-r001/2
```

The corr

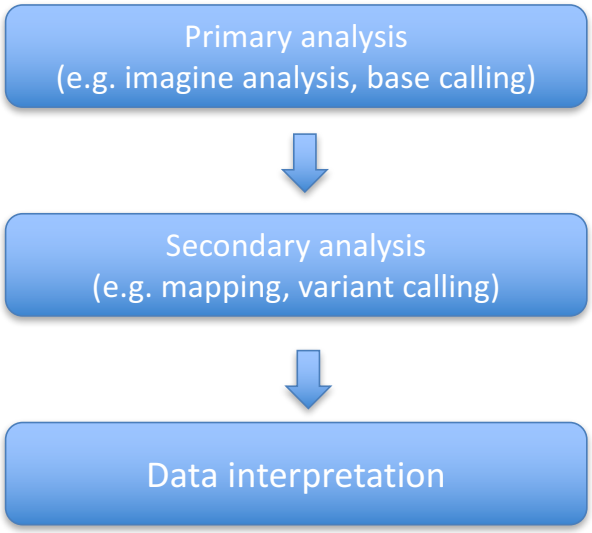
Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next segment
8	PNEXT	Int	$[0,2^{29}-1]$	Position of the mate/next segment
9	TLEN	Int	$[-2^{29}+1,2^{29}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

```
@HD VN:1.
@SQ SN:re
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

samtools.sourceforge.net/SAM1.pdf

}

Understanding all analysis steps

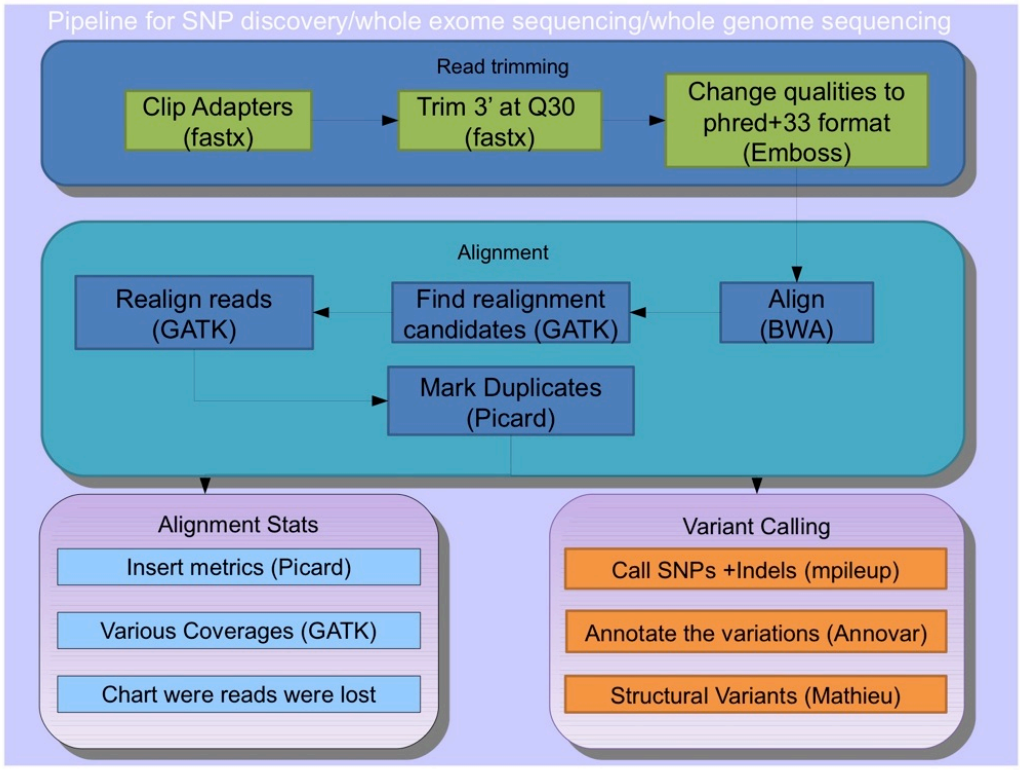


McGill

No perfect recipe

- Analysis steps, tools, parameters… Everything depends on your specific project
- You may want to use stringent cutoffs… Or loose cutoffs… It depends on what you will be doing with the data

Exome processing pipeline



Louis Letourneau, Innovation Center

McGill

Pipeline Outline

- Exome and whole-genome sequencing
  - Pre-processing (remove adapters, trimming, …)
  - Mapping
    - BWA
    - Mosaik, …
  - Variant calling
    - SNP
    - Indels
    - Copy number
    - Structural variants
  - Variant annotation

Pipeline Outline

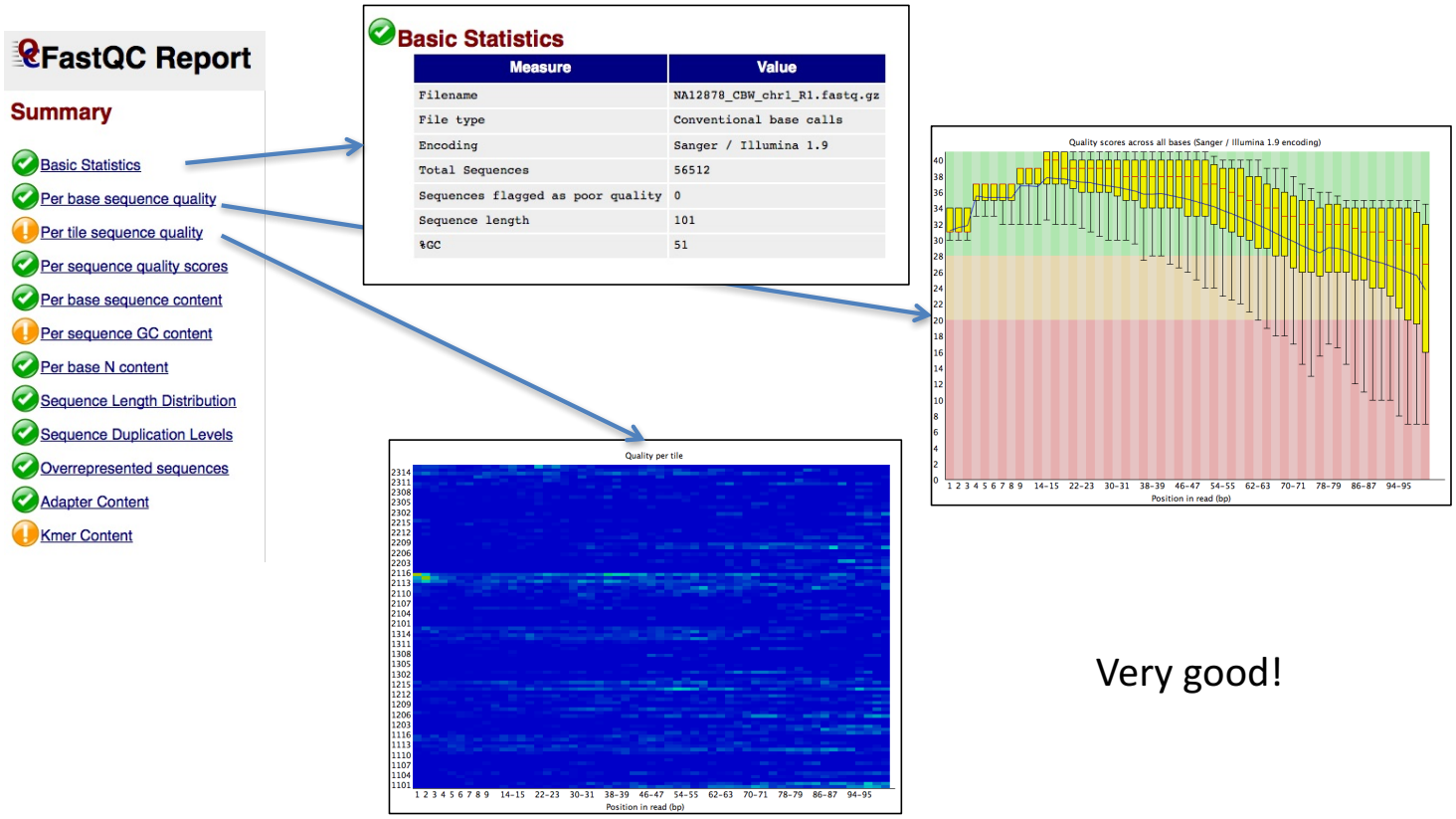
- Exome and whole-genome sequencing
  - Pre-processing (remove adapters, trimming, …)
  - Mapping
    - BWA
    - Mosaik, …
  - Variant calling
    - SNP
    - Indels
    - Copy number
    - Structural variants
  - Variant annotation

Importance of quality control

- Before you start an analysis, it’s very important to look at your raw data!
- Are all of your samples sequenced using the same protocol and instruments?
- Are there any technical issues affecting some of the samples?
- This is especially important if you plan to compare different samples or different conditions

McGill

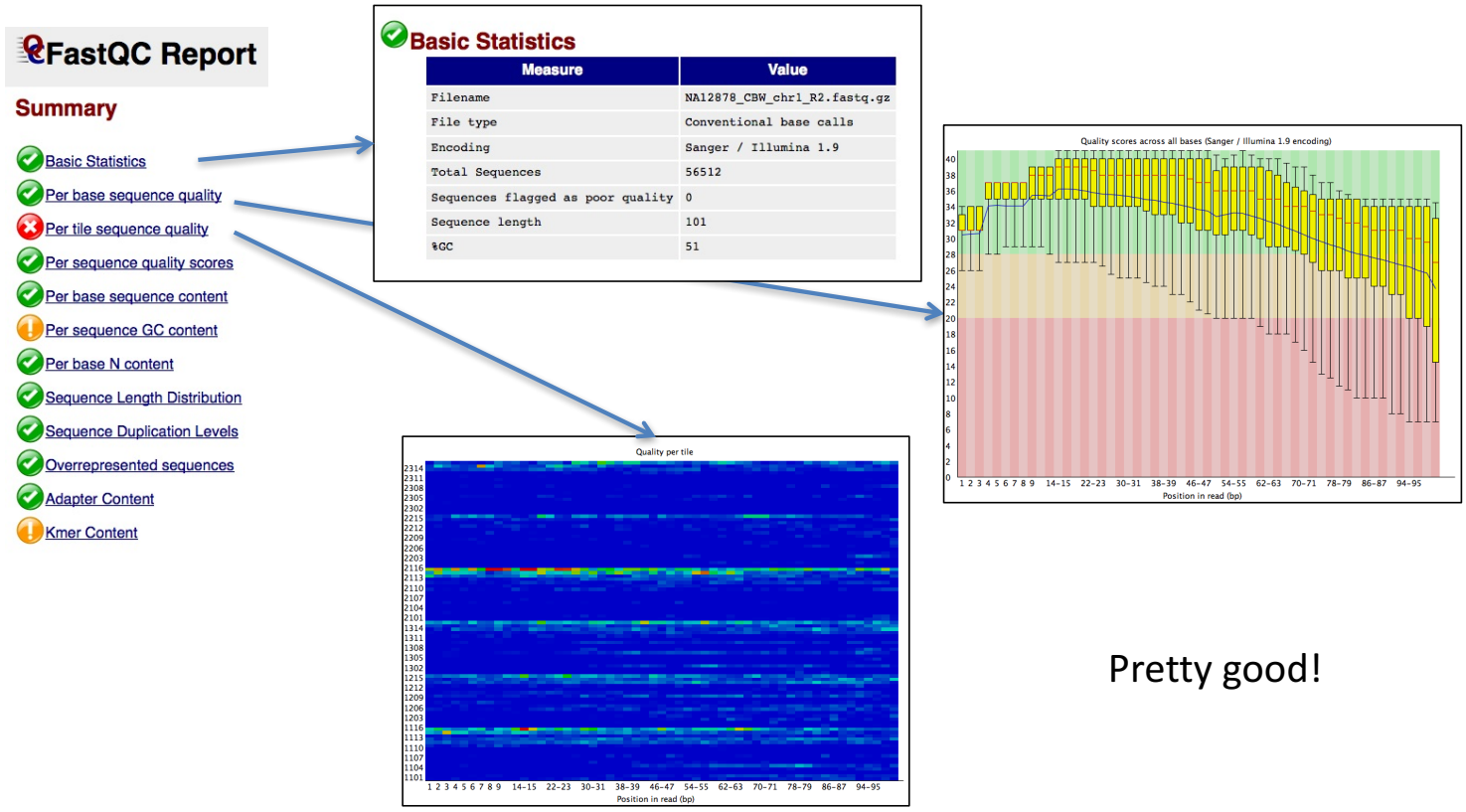
Running FastQC on read 1



Very good!

McGill

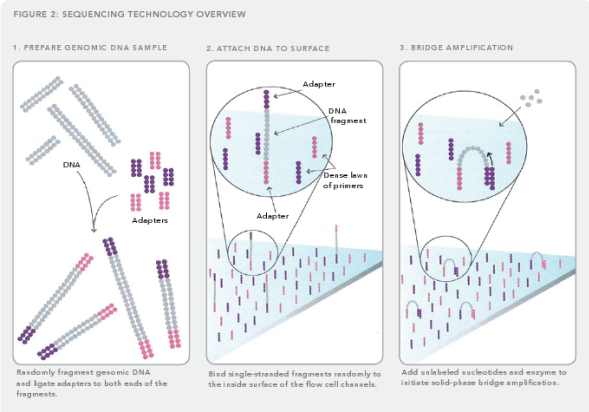
Running FastQC on read 2



Pretty good!

McGill

Adapters sequences in reads

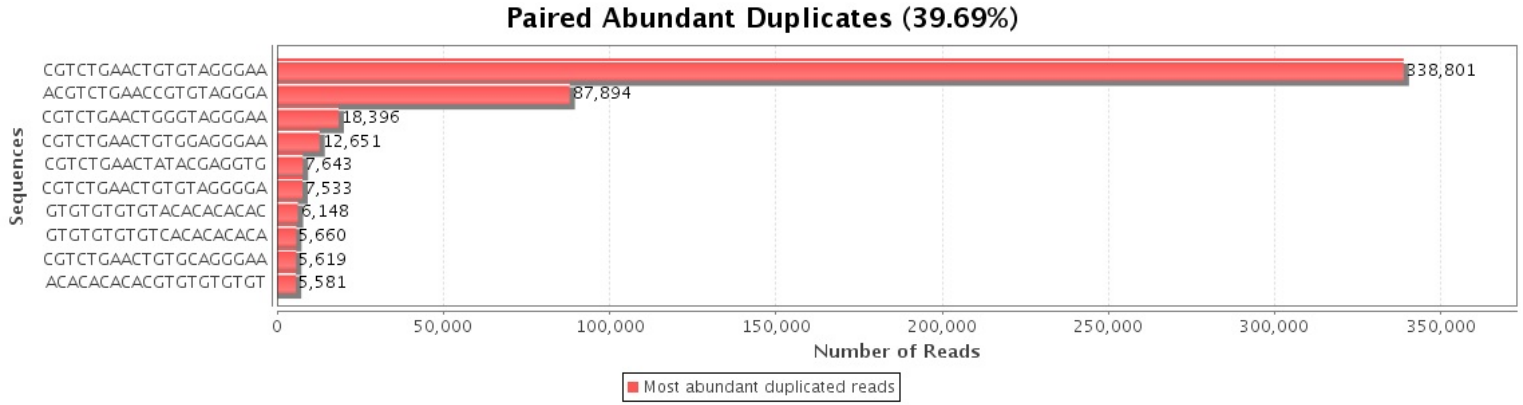


http://www.illumina.com

http://srna-workbench.cmp.uea.ac.uk

McGill

Check for over-represented sequences



McGill

Read trimming tools

For example, Trimmomatic performs a variety of useful trimming tasks for illumina paired-end and single ended data:

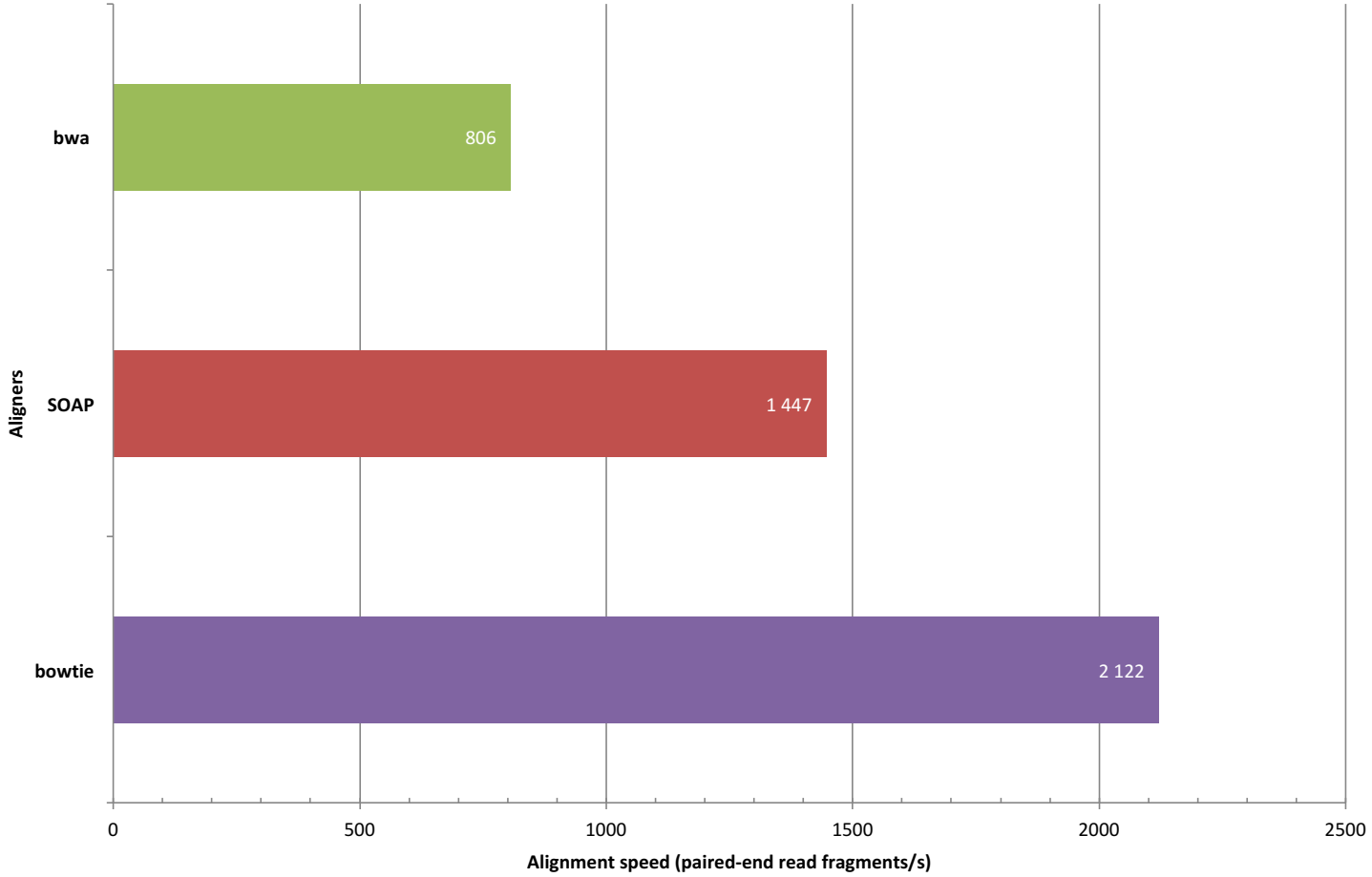
- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- CROP: Cut the read to a specified length
- HEADCROP: Cut the specified number of bases from the start of the read
- MINLEN: Drop the read if it is below a specified length
- TOPHRED33: Convert quality scores to Phred-33
- TOPHRED64: Convert quality scores to Phred-64

Pipeline Outline

- Exome and whole-genome sequencing
  - Pre-processing (remove adapters, trimming, …)
  - Mapping
    - BWA
    - Mosaik, …
  - Variant calling
    - SNP
    - Indels
    - Copy number
    - Structural variants
  - Variant annotation

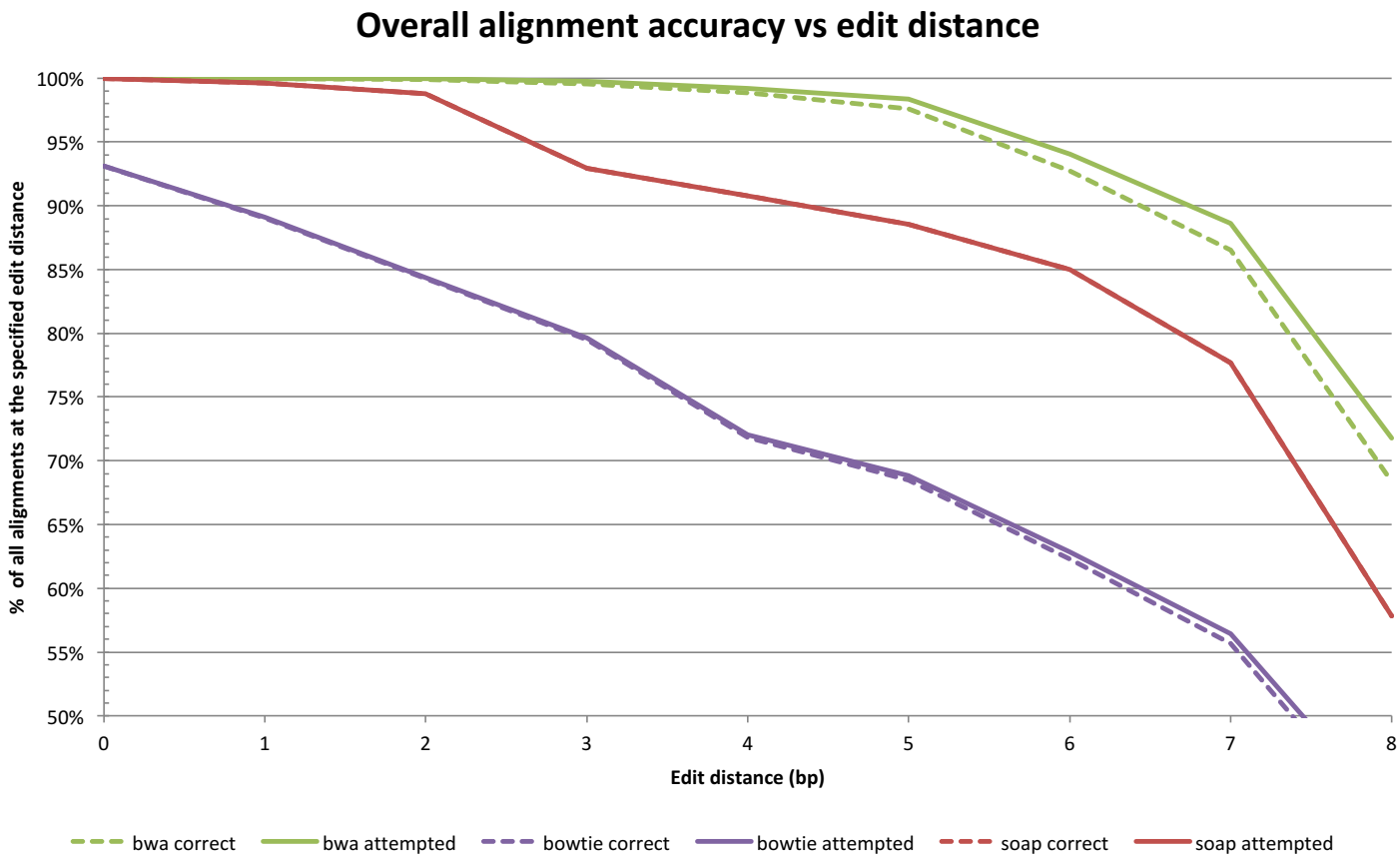
	Illumina	AB SOLiD	Roche 454	Helicos	gapped	all alignments	multithreaded
BFAST	X	X	X	X	X	X	X
Bowtie	X	X				X	X
BWA	X				X	X	X
Corona Lite		X				X	
ELAND	X						
GenomeMapper	X				X	X	X
gnumap	X				X	X	X
karma	X	X	X		*		
MAQ	X	X					
MOSAIK	X	X	X	X	X	X	X
MrFAST	X				X	X	
MrsFAST	X					X	
Novoalign	X				X	X	*
RMAP	X					X	
SeqMap	X				X	X	
SHRiMP	X	X	X	X	X	X	
Slider	X					X	
SOAP2	X					X	X
SSAHA2	X		X		X	X	
SOCS		X					X
SXOligoSearch	X		X		X	X	
Zoom	X	X			*	X	

Alignment Speed (30M 100bp PE reads aligned to hg19 using 1 core)



Michael Stromberg, bioinformatics.ca

McGill



- The **solid lines** denote what percentage of reads at a given edit distance were aligned. The **dashed lines** denote the percentage which were aligned correctly.

Short read mappers

- Trade-off between speed and accuracy
- For DNA data, BWA performs well and is frequently viewed as the default mapper
- It’s important to also do quality recalibration for better SNV calling (how different are predicted vs observed qualities)
- Mapping for RNA-Seq is different (see tomorrow)

Pipeline Outline

- Exome and whole-genome sequencing
  - Pre-processing (remove adapters, trimming, …)
  - Mapping
    - BWA
    - Mosaik, …
  - Variant calling
    - SNP
    - Indels
    - Copy number
    - Structural variants
  - Variant annotation

SNP Discovery: Goal

Michael Stromberg, bioinformatics.ca

McGill

SNP Discovery: Base Qualities

SNP and indel calling tools

Table 2 | **Computational tools for cancer genomics**

Category	Method	URL	Comments	Refs
Alignment	MAQ	http://maq.sourceforge.net	Used by most cancer genome papers so far	108
	BWA	http://bio-bwa.sourceforge.net	Replacing MAQ. Considerably faster	109
	ELAND	http://www.illumina.com		117
	SSAHA2	http://www.sanger.ac.uk/resources/software/ssaha2	Used to validate location of reads	39,110
	Bowtie	http://bowtie-bio.sourceforge.net/index.shtml		111
	SOAP2	http://soap.genomics.org.cn		112
	SHRiMP	http://compbio.cs.toronto.edu/shrimp		113
	Corona Lite	http://solidsoftwaretools.com/gf/project/corona	Used for SOLiD	
	BFAST	http://bfast.sourceforge.net	Mainly used for SOLiD	114
Mutation calling	SNVMix	http://www.bcgsc.ca/platform/bioinfo/software/SNVMix		80
	CASAVA	http://www.illumina.com/software/genome_analyzer_software.ilmn		117
	Samtools	http://samtools.sourceforge.net		104
	Unified genotyper	http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper		107
	VarScan	http://varscan.sourceforge.net		105
Indel calling	Pindel	http://www.ebi.ac.uk/~kye/pindel		106
Copy number analysis	CBS	https://r-forge.r-project.org/R/?group_id=702	CBS used on tumour/normal ratios calculated in fixed windows	115
	SegSeq	http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=182		26
Prediction of mutation functional effect	SIFT	http://blocks.fhcrc.org/sift/SIFT.html http://sift.jcvi.org		81
	Polyphen-2	http://genetics.bwh.harvard.edu/pph2		83
	XVAR	http://xvar.org		119
	CHASM			85
Visualization	CIRCOS	http://mkweb.bcgsc.ca/circos	Essentially all papers use CIRCOS to display genomic events	120
	IGV	http://www.broadinstitute.org/igv	IGV is used to display genomic events and for manual review	

A list of additional alignment methods with a brief description of each is constantly updated at http://en.wikipedia.org/wiki/List_of_sequence_alignment_software.

Meyerson et al. 2010

SNP and Indel calling is a large-scale Bayesian modeling problem

Bayesian model

$$\Pr\{G|D\} = \frac{\Pr\{G\}\Pr\{D|G\}}{\Sigma_i \Pr\{G_i\}\Pr\{D|G_i\}}, \text{ [Bayes' rule]}$$

Prior of the genotype

Likelihood of the genotype

$$\Pr\{D|G\} = \prod_j \left(\frac{\Pr\{D_j|H_1\}}{2} + \frac{\Pr\{D_j|H_2\}}{2} \right) \text{ where } G = H_1 H_2$$

Diploid assumption

$$\Pr\{D|H\} \text{ is the haploid likelihood function}$$

- Inference: what is the genotype G of each sample given read data D for each sample?
- Calculate via Bayes' rule the probability of each possible G
- Product expansion assumes reads are independent
- Relies on a likelihood function to estimate probability of sample data given proposed haplotype

www.broadinstitute.org

McGill

SNP genotype likelihoods

$$\Pr\{D_j|H\} = \Pr\{D_j|b\},\ [\text{single base pileup}]$$

$$\Pr\{D_j|b\} = \begin{cases} 1-\epsilon_j & D_j = b, \\ \epsilon_j & \text{otherwise.} \end{cases}$$

- All diploid genotypes (AA, AC, …, GT, TT) considered at each base
- Likelihood of genotype computed using only pileup of bases and associated quality scores at given locus
- Only “good bases” are included: those satisfying minimum base quality, mapping read quality, pair mapping quality, NQS

Strategies that improve variant calling

- Local realignment
- Duplicate marking
- Base quality recalibration
- Population structure and imputation

McGill

Strategies that improve variant calling

- Local realignment
- Duplicate marking
- Base quality recalibration
- Population structure and imputation

McGill

Local realignement

Before realignement

After realignement

DePristo et al. *Nat Genet* 2011

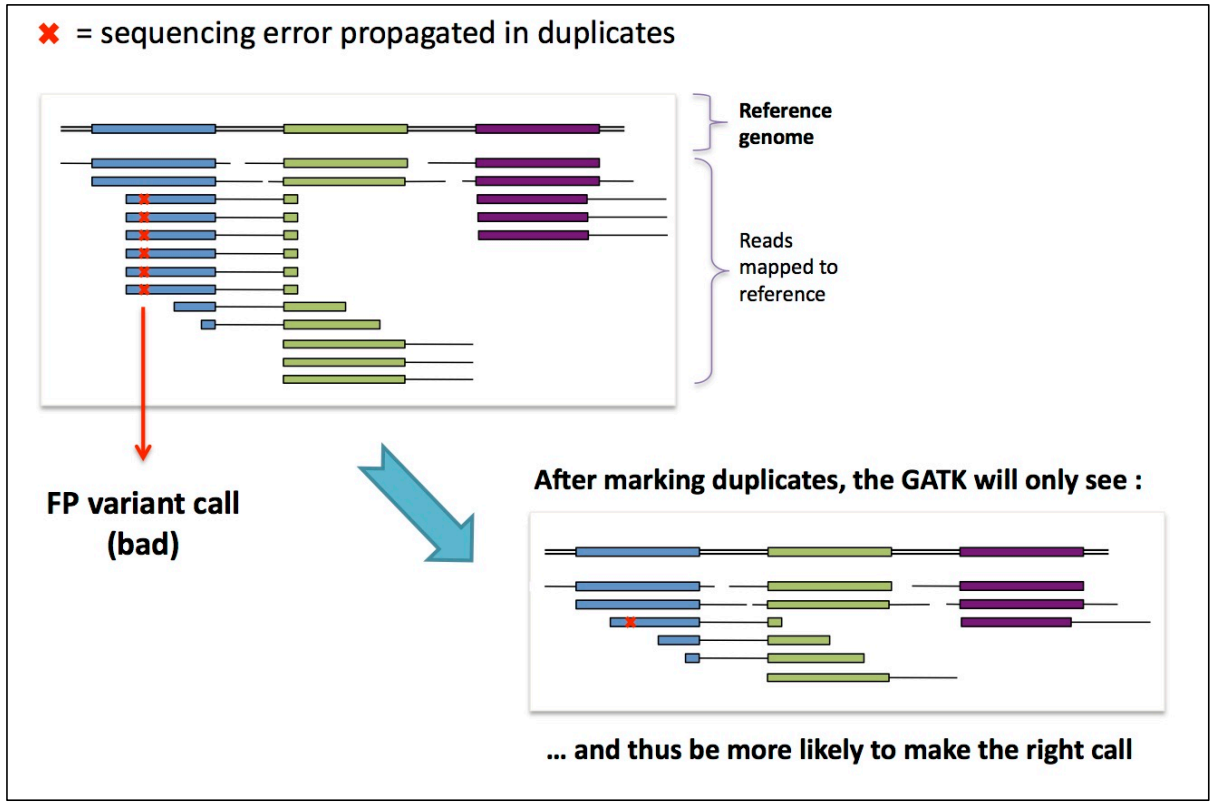
McGill

Strategies that improve variant calling

- Local realignment
- Duplicate marking
- Base quality recalibration
- Population structure and imputation

McGill

Duplicate marking



www.broadinstitute.org

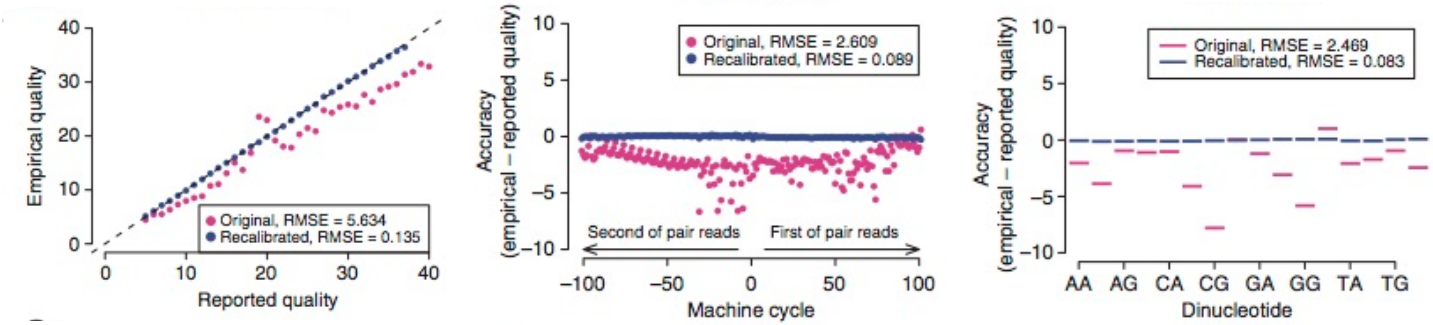
McGill

Strategies that improve variant calling

- Local realignment
- Duplicate marking
- Base quality recalibration
- Population structure and imputation

McGill

Base quality recalibration



Adapted from DePristo et al. *Nat Genet* 2011

McGill

Strategies that improve variant calling

- Local realignment
- Duplicate marking
- Base quality recalibration
- Population structure and imputation

McGill

Using haplotypes for base calling

- Suppose that only 2 haplotypes have been observed in a population:

```
Chr1: ..........A....T.......G..........
Chr1: ..........C....G.......A..........
```

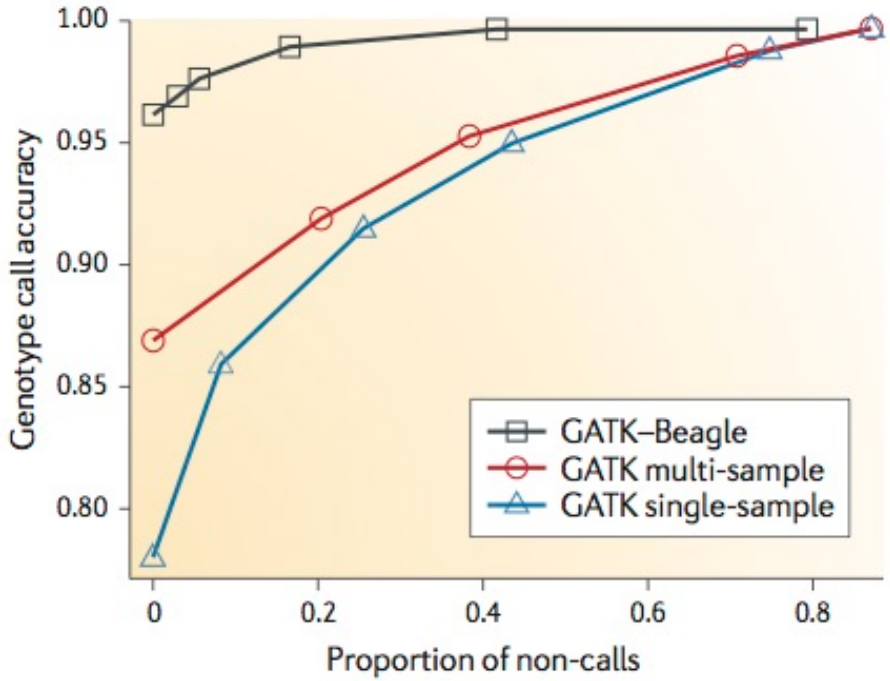
- And that you observe the following reads:

```
......A....N.......G..
   ..A....N.......G.....
  ...A....N.......G...
```

- Can you guess the value of `N` ?

McGill

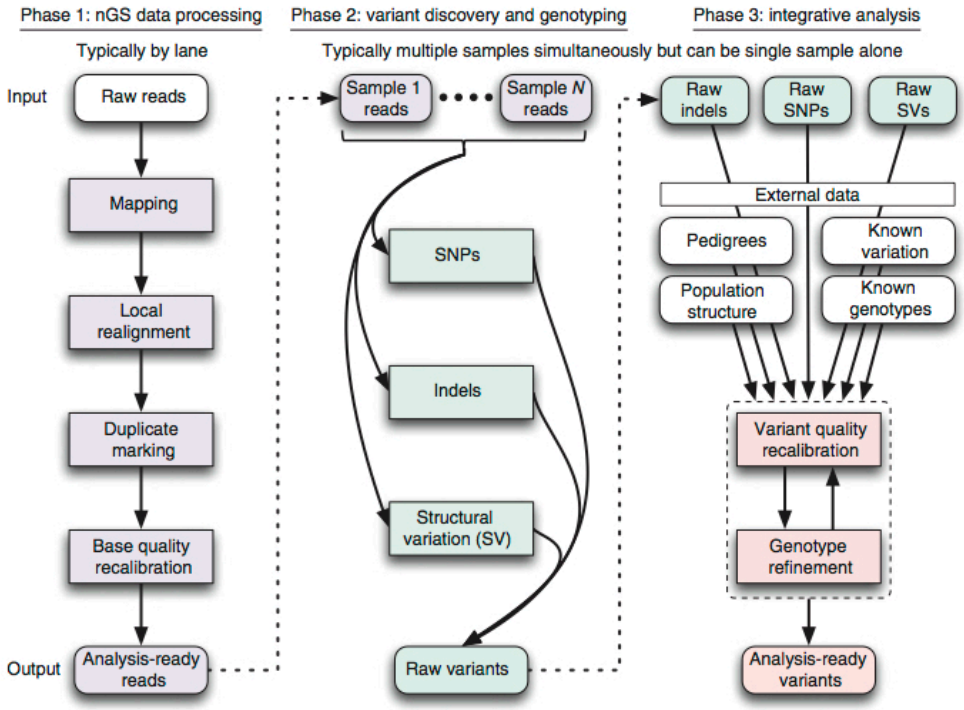
Impact of using multi-samples and haplotype information



Nielsen et al. *Nat Rev Genet* 2011

McGill

GATK framework



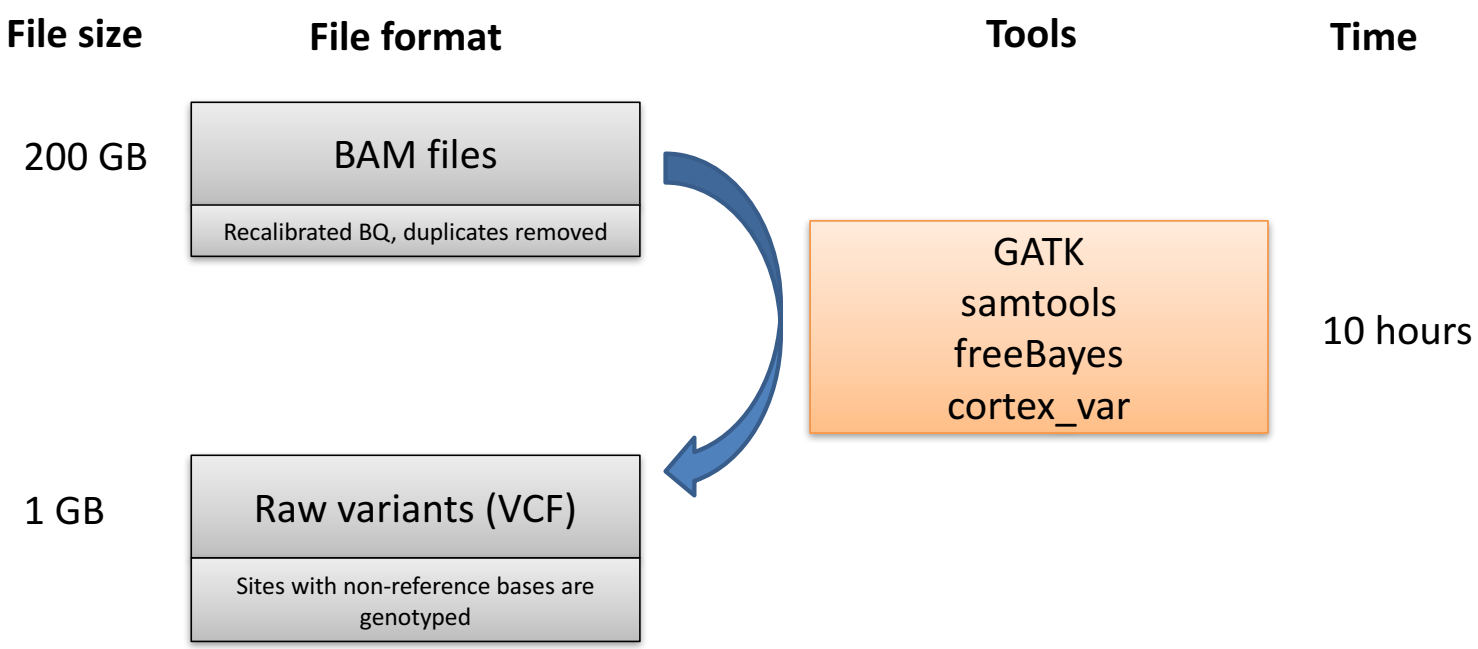
DePristo et al. *Nat Genet* 2011

McGill

File size	File format	Tools	Time
200 GB	BAM files Recalibrated BQ, duplicates removed		

Adapted from Mark DePristo

McGill



Adapted from Mark DePristo

McGill

Pipeline Outline

- Exome and whole-genome sequencing
  - Pre-processing (remove adapters, trimming, …)
  - Mapping
    - BWA
    - Mosaik, …
  - Variant calling
    - SNP
    - Indels
    - Copy number
    - Structural variants
  - Variant annotation

Structural Variants (SVs)

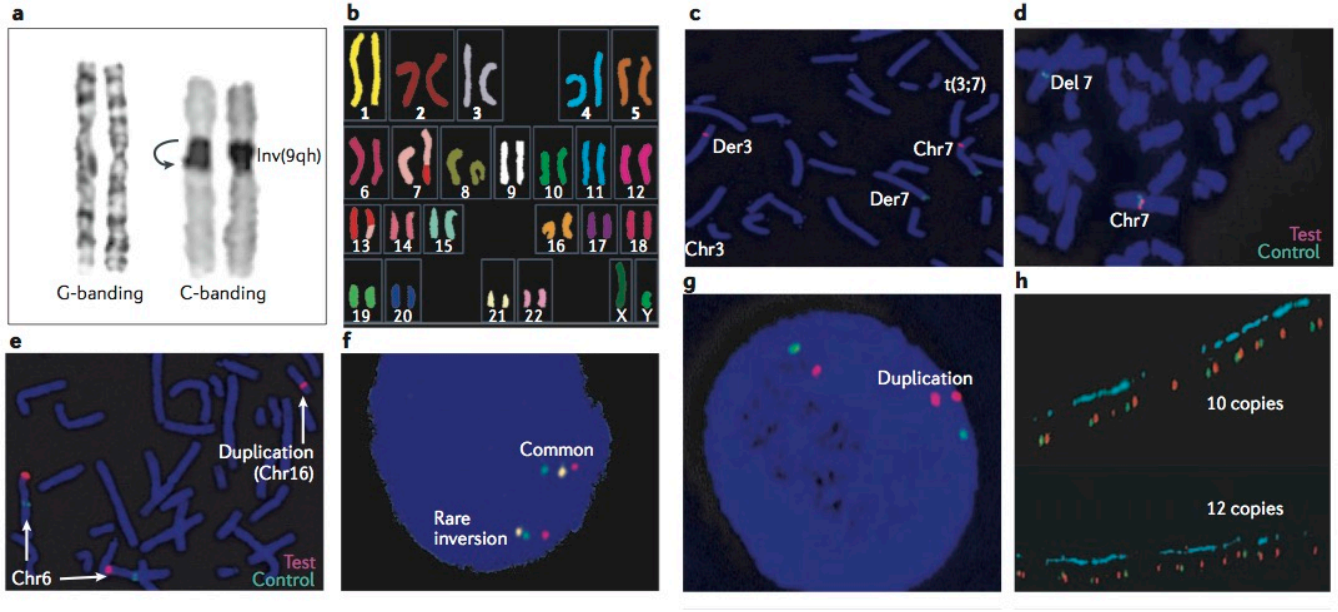
Structural Variants (SVs): Genomic rearrangements that affect >50bp (or 100bp, or 1Kb) of sequence, including:

- deletions
- novel insertions
- inversions
- mobile-element transpositions
- duplications
- translocations

Adapted from Alkan et al. *Nat Rev Genet* 2011

McGill

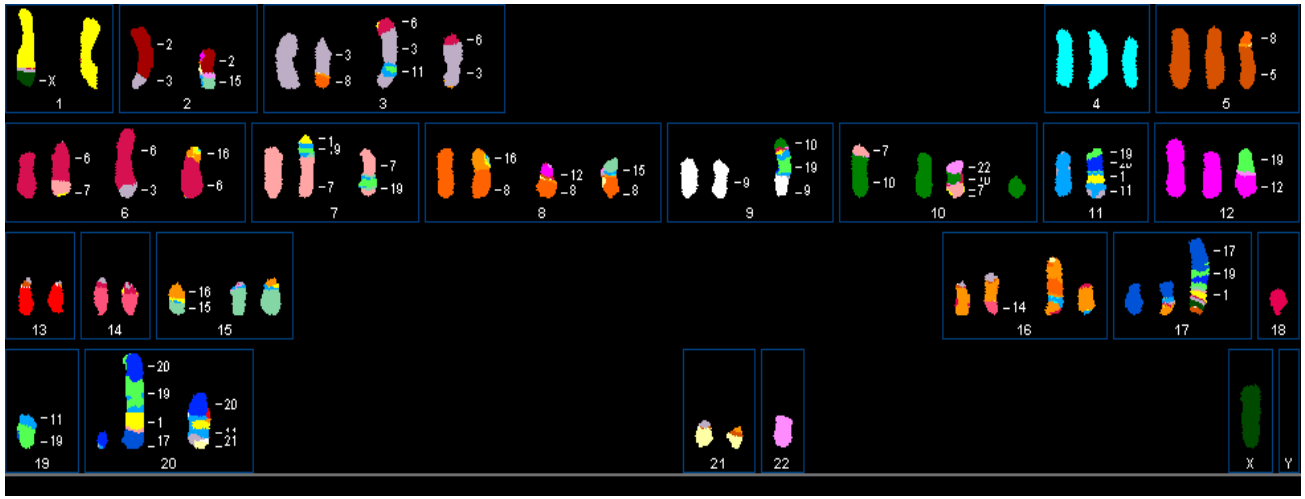
Detection and confirmation of SVs



Feuk et al. *Nat Rev Genet* 2006

McGill

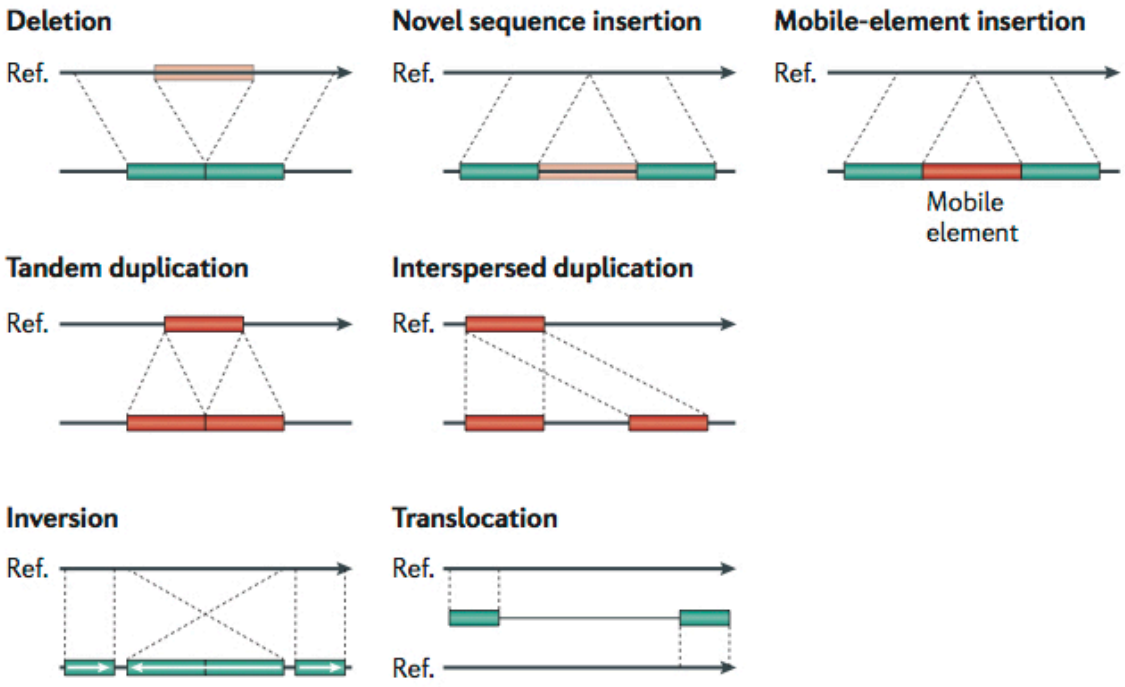
Structural variants in cancer



Can higher resolution maps help identify recurrent aberrations and driver mutations in cancer?

McGill

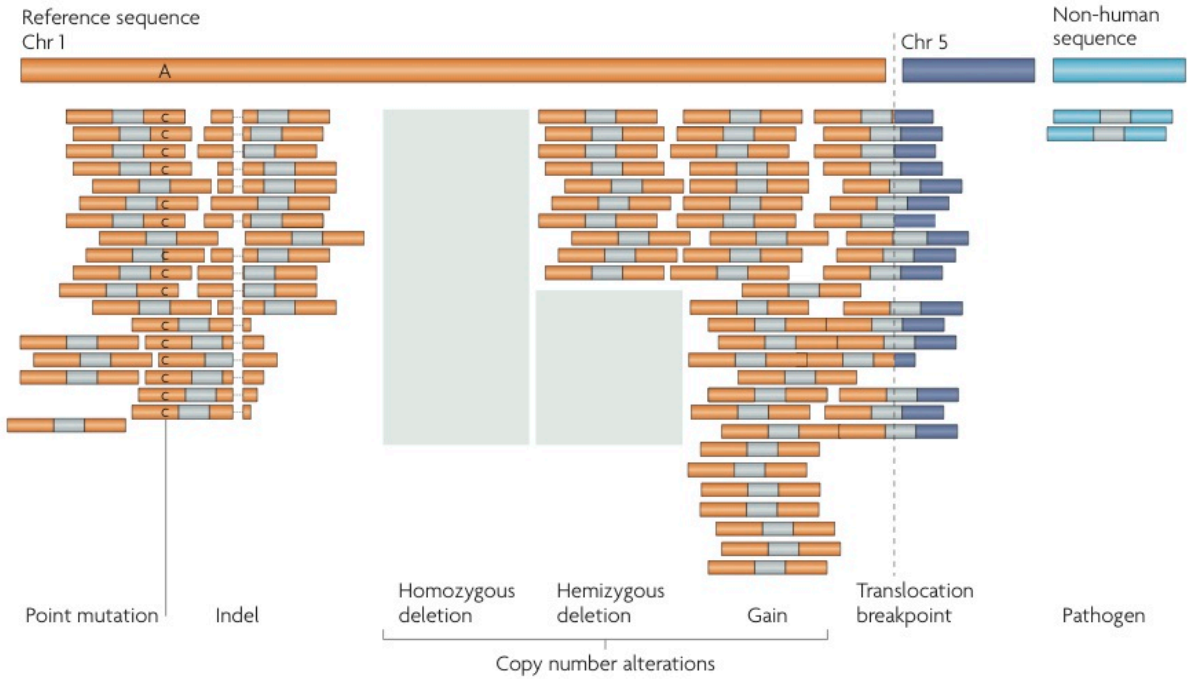
Classes of SVs



Alkan et al. *Nat Rev Genet* 2011

McGill

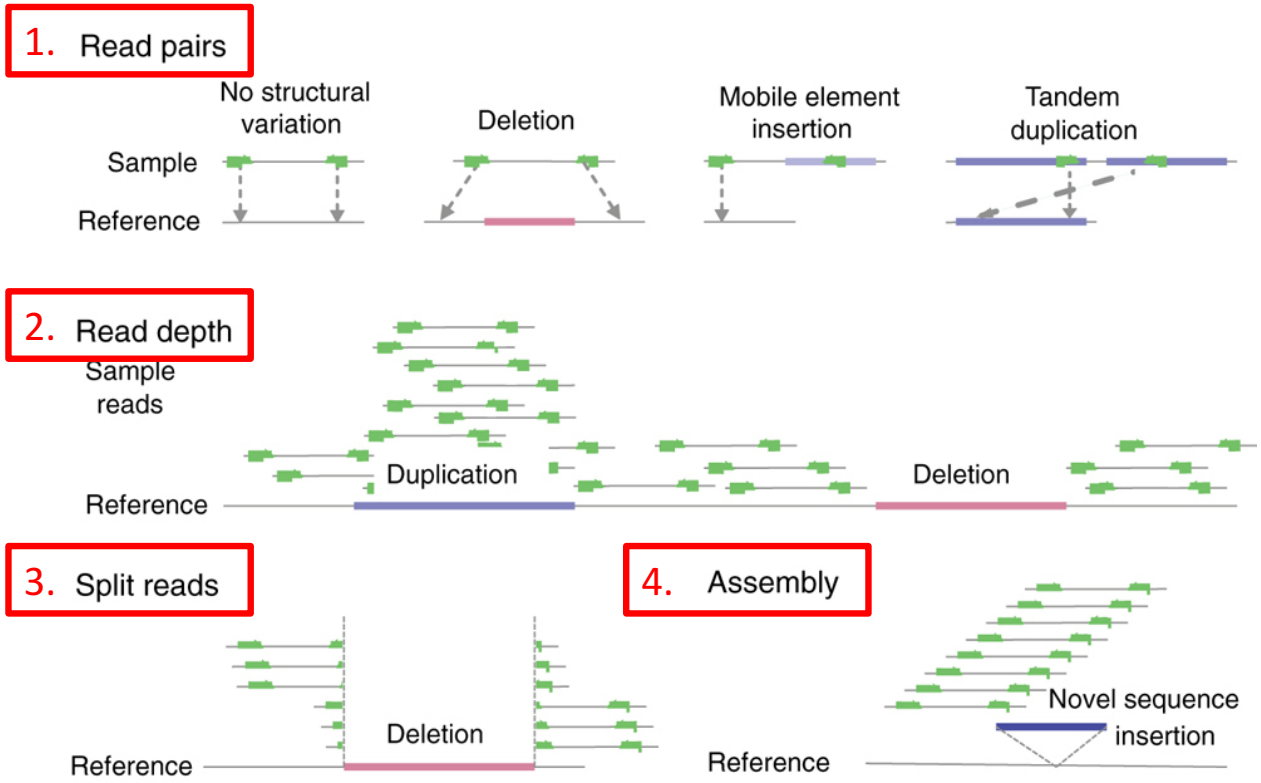
Detecting SVs from NGS data



Meyerson et al. *Nat Rev Genet* 2010

McGill

Strategies for calling SVs from NGS data

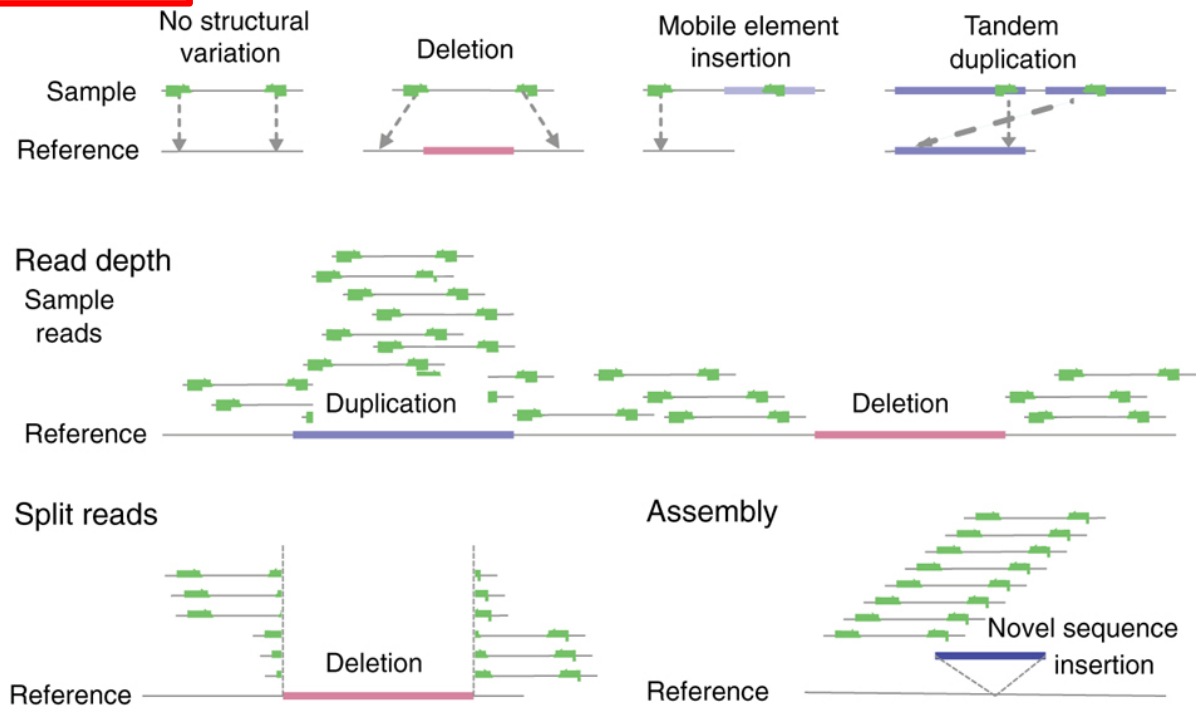


Baker *Nat Methods* 2012

McGill

Strategies for calling SVs from NGS data

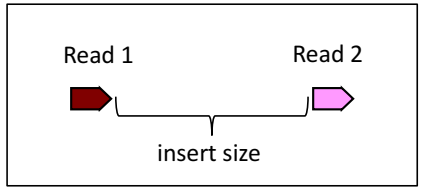
1. Read pairs

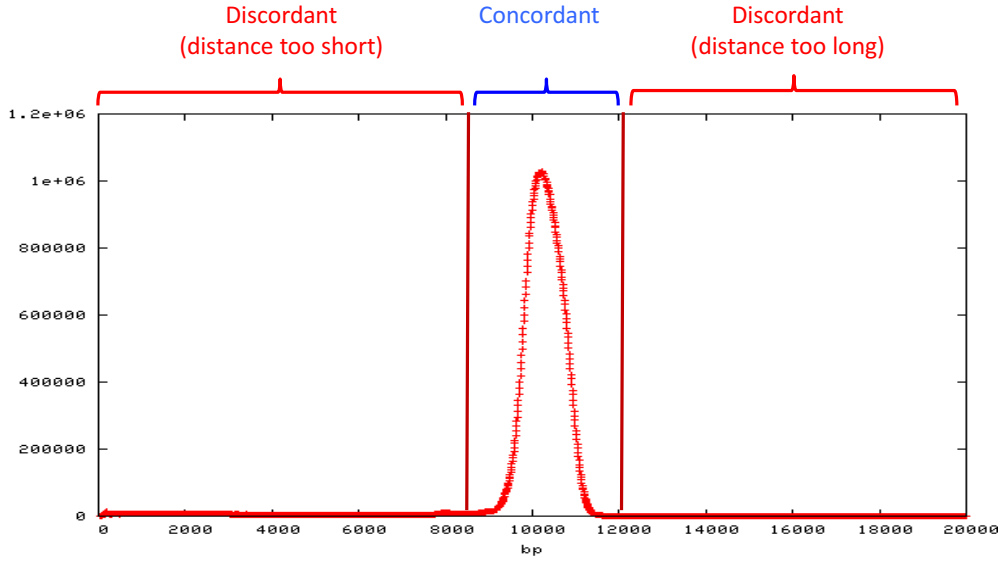


Baker *Nat Methods* 2012

McGill

Discordant read pairs



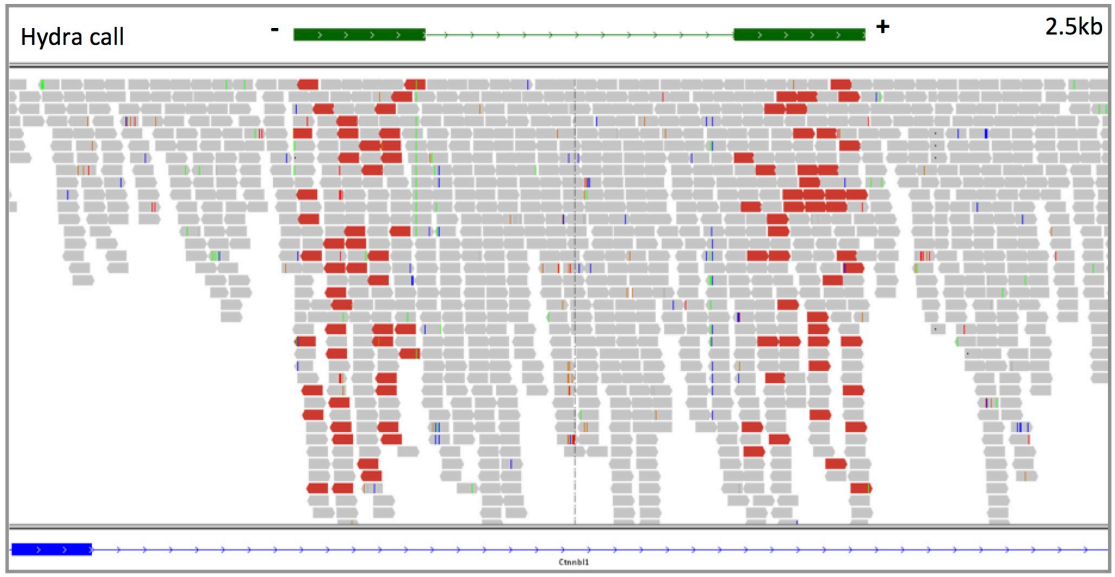


Genomic distance between mapped paired tags

Reads pairs are also Discordant when order or orientation isn't as expected.

McGill

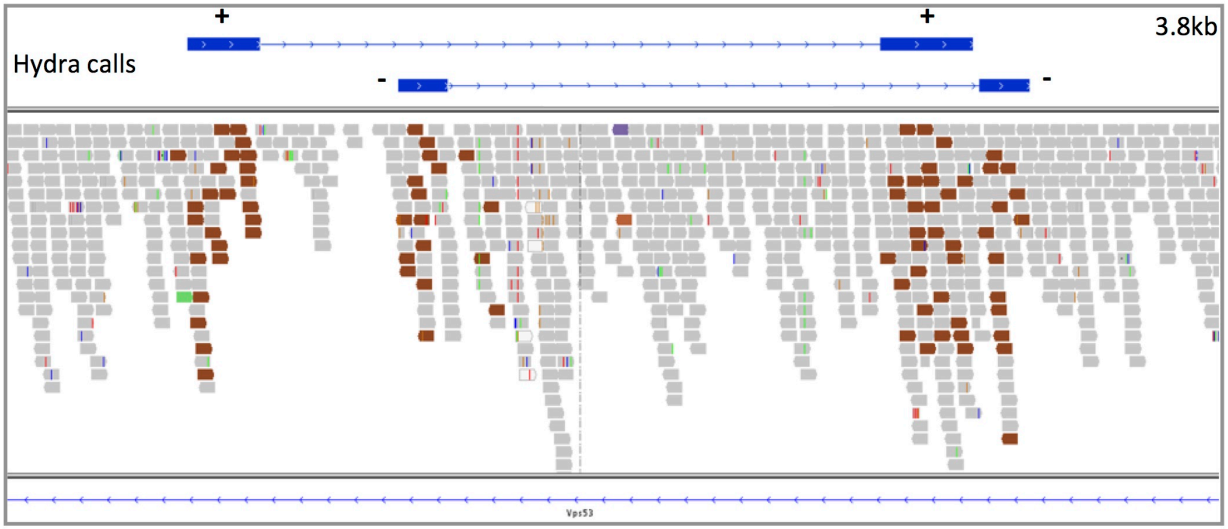
Visual validation: a duplication



Aaron Quinlan

McGill

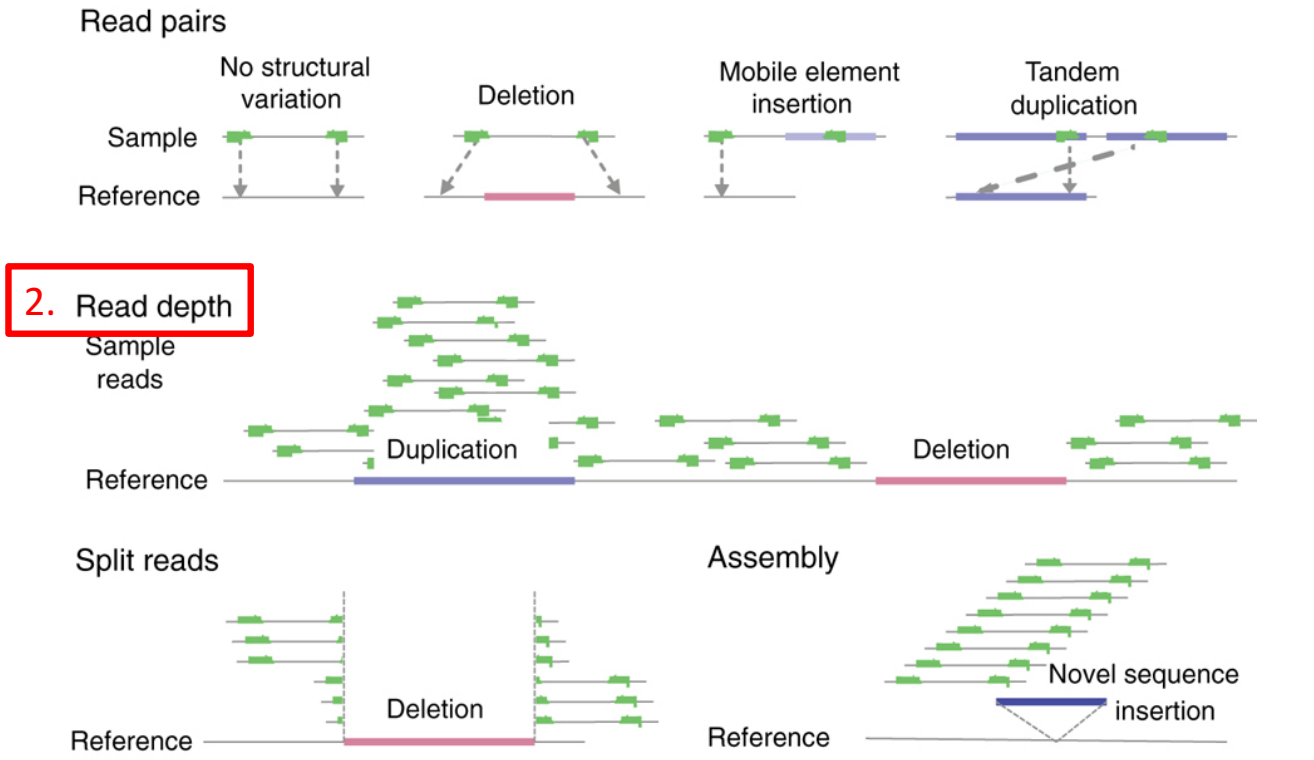
Visual validation: an inversion



Aaron Quinlan

McGill

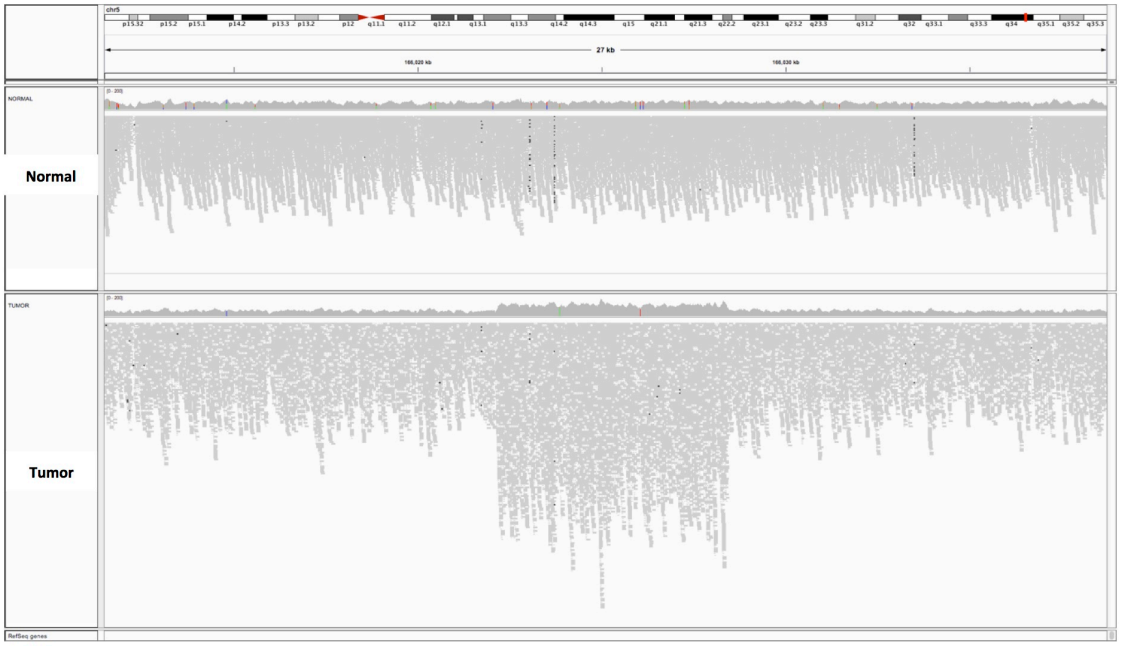
Strategies for calling SVs from NGS data



Baker *Nat Methods* 2012

McGill

Read-depth



Basic approach:

1) Count reads in sliding windows (e.g., 5kb) across the genome.

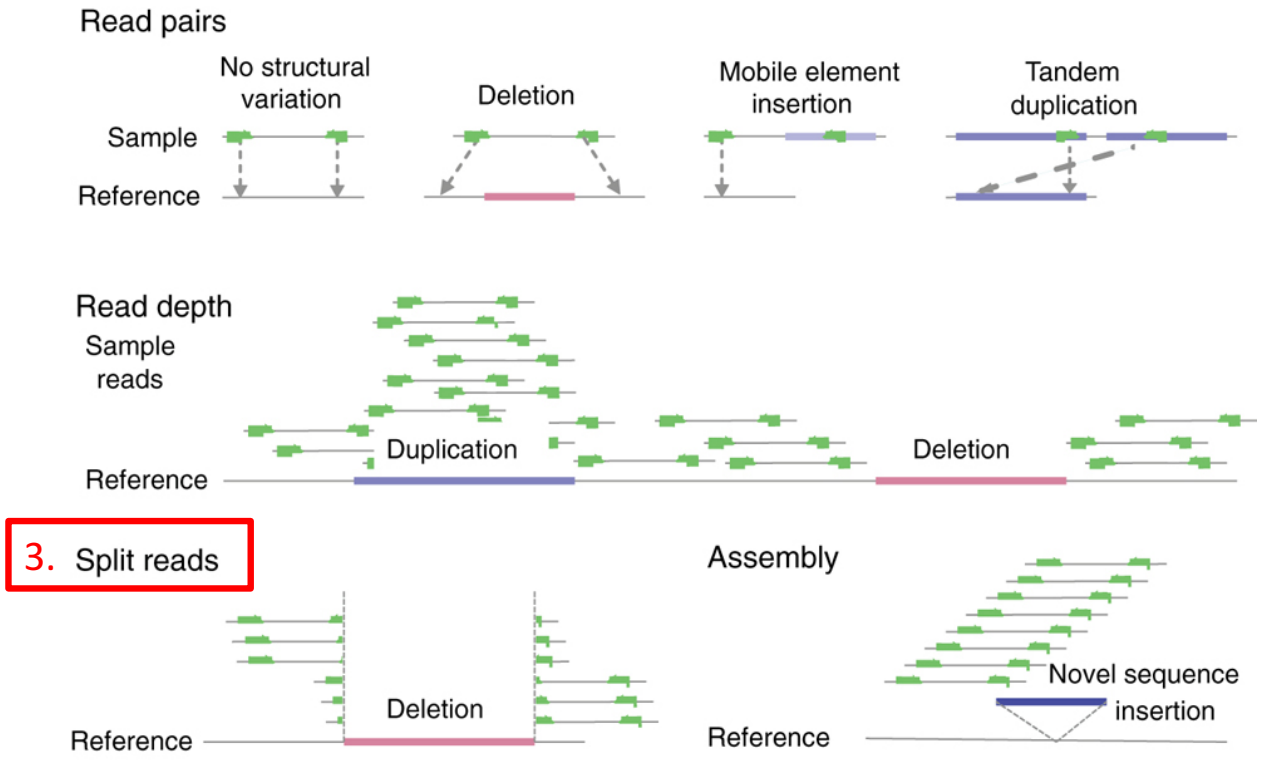
2) Normalize for GC bias.

3) Use segmentation to define CNAs (similar to array-CGH data).

Aaron Quinlan

McGill

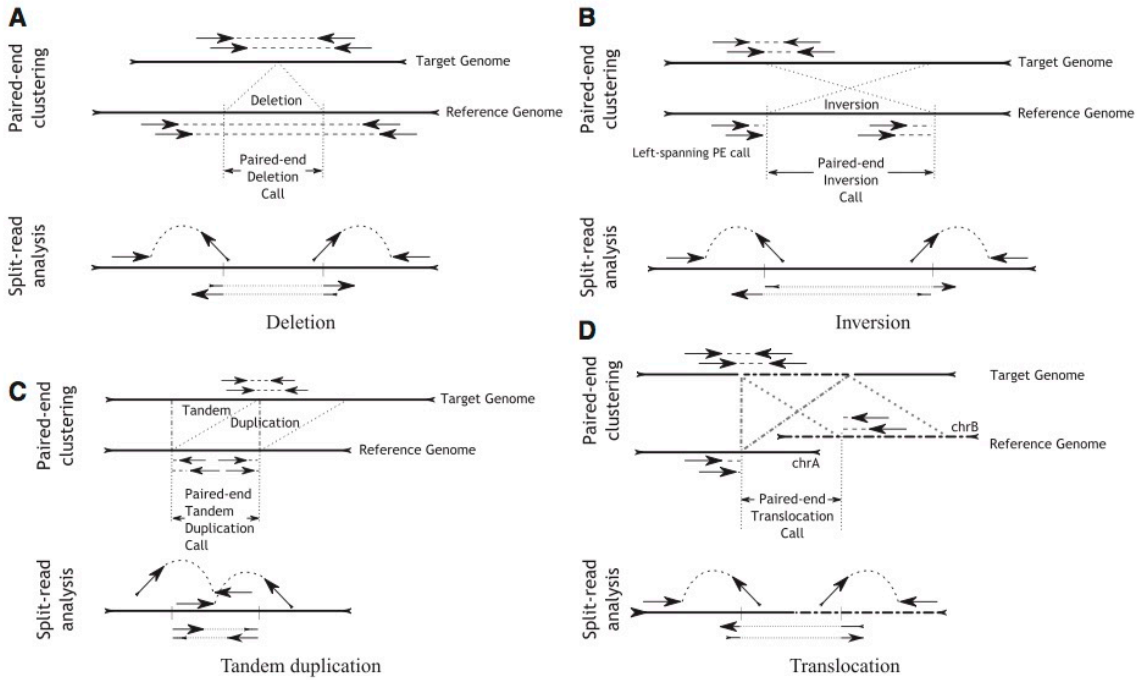
Strategies for calling SVs from NGS data



Baker *Nat Methods* 2012

McGill

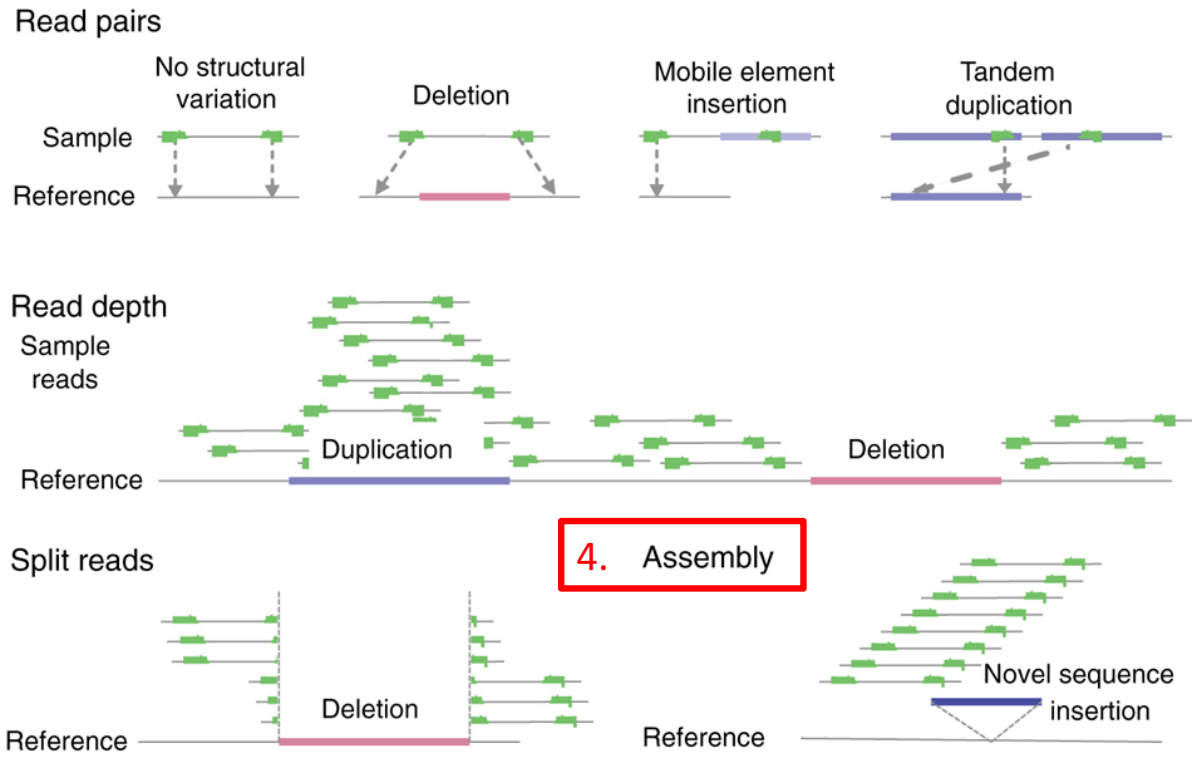
Split reads



Rausch et al. *Bioinformatics* 2012

McGill

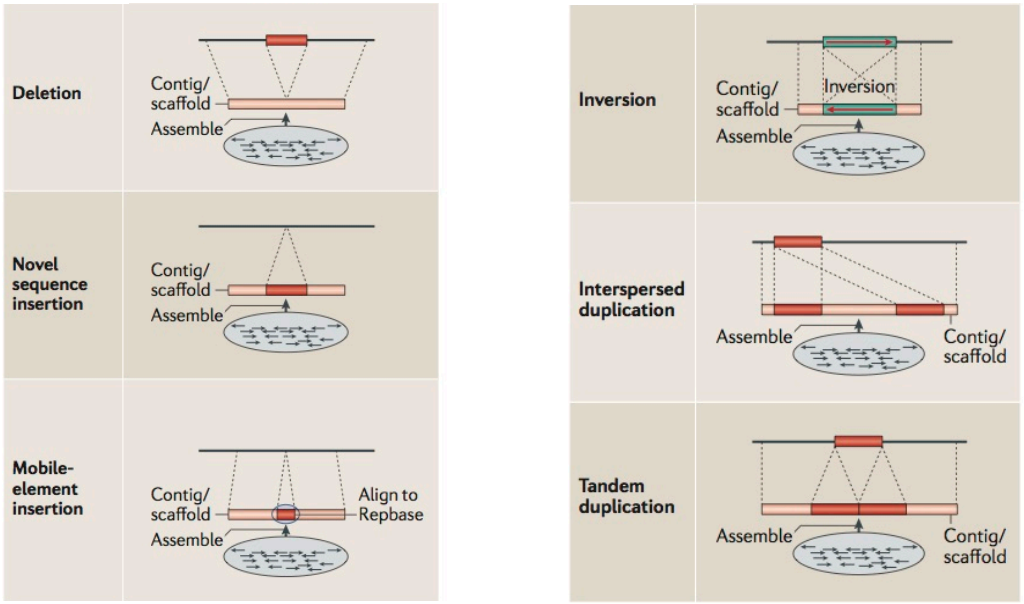
Strategies for calling SVs from NGS data



Baker *Nat Methods* 2012

McGill

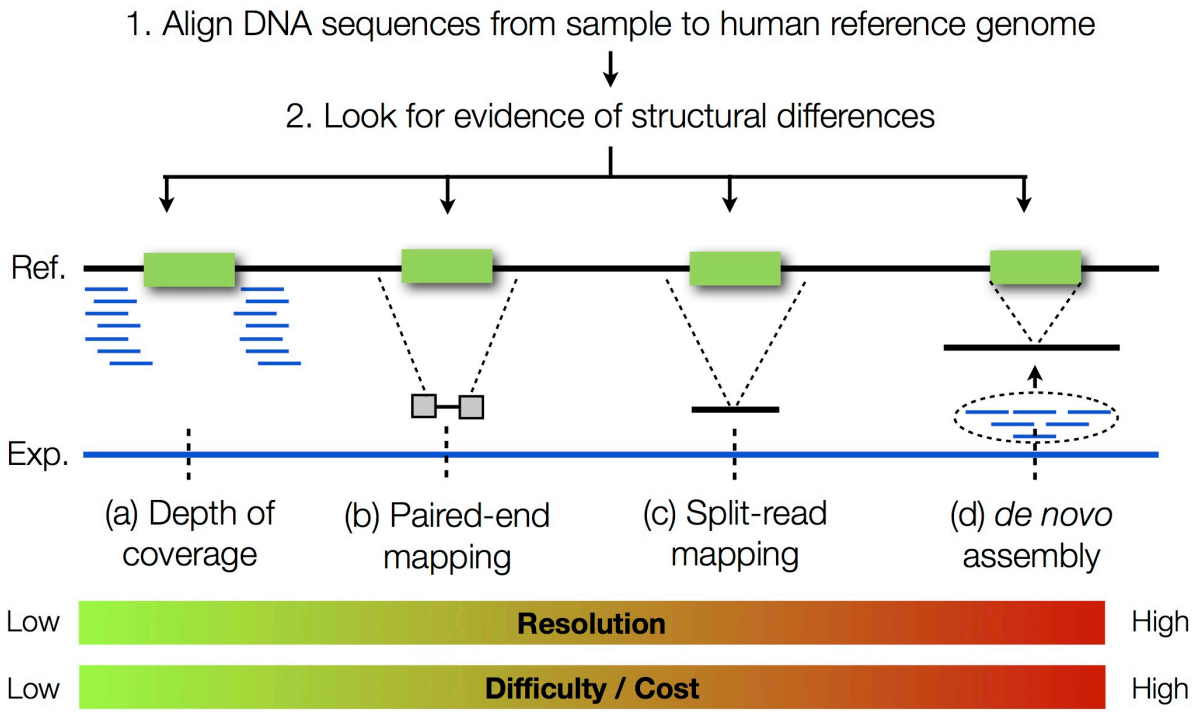
De novo assembly for SVs



Adapted from Alkan et al. *Nat Rev Genet* 2011

McGill

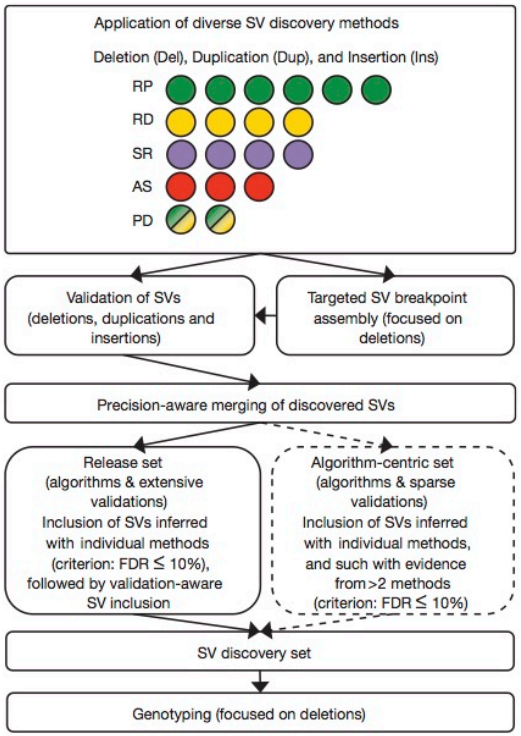
Summary of strategies for calling SVs



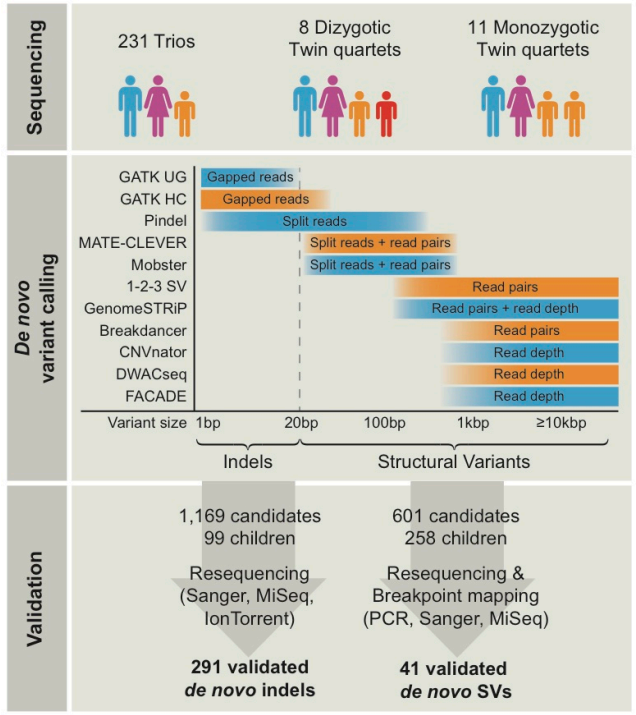
Aaron Quinlan

McGill

Bottom line: try many methods and validate



Mills et al. *Nature* 2011



Kloosterman et al. *Genome Res* 2015

McGill

Pipeline Outline

- Exome and whole-genome sequencing
  - Pre-processing (remove adapters, trimming, …)
  - Mapping
    - BWA
    - Mosaik, …
  - Variant calling
    - SNP
    - Indels
    - Copy number
    - Structural variants
  - Variant annotation

VCF format

```
##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">
#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
```

Mandatory header line

Mandatory header line

Reference base

Alternative base

Quality score

Allele frequency, read depth, etc.

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

McGill

Variant filtering

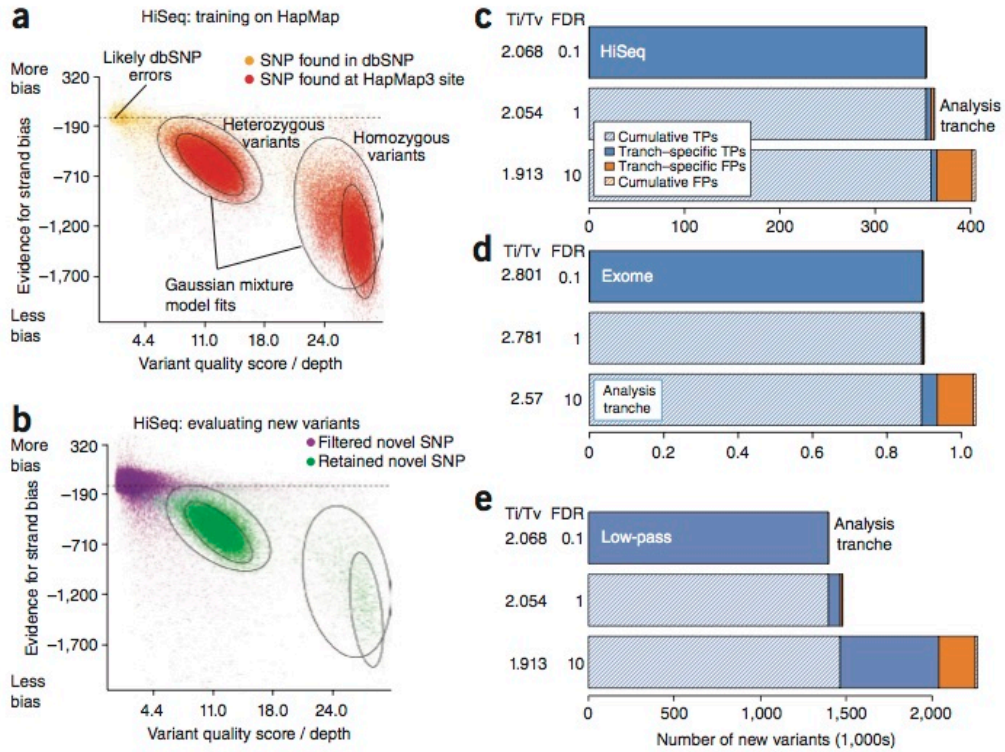
Raw variant calls have a lot of false positives. How to filter?

1. Manual filtering based on different parameters (e.g. using GATK VariantFiltration or snpSift):
   - Based on quality score, depth of coverage, etc.
   - Difficult and requires time and expertise
2. Learn the filters from the data itself (e.g. GATK VariantRecalibrator):
   - Better rank-order variants based on their likelihood of being real

QC: HapMap & dbSNP

- **International HapMap Project** (phase III)
  - 1301 individuals in 11 populations genotyped
  - ~1 SNP per 2 kb
  - Proxy for **false negatives**
- **dbSNP** (build 130)
  - 14 million SNPs in human genome
  - Varying quality
  - Proxy for **false positives**

Variant Quality Recalibration



DePristo et al. *Nat Genet* 2011

McGill

Variant Annotation

- Variant and mutation databases (dbSNP, DGV, COSMIC, etc.).
- Functional impact for coding variants (synonymous, non-synonymous, etc.)
- Sequence conservation and other genome annotation features
- Other functional data for non-coding variants (e.g. ENCODE)

- What to do for structural variants?

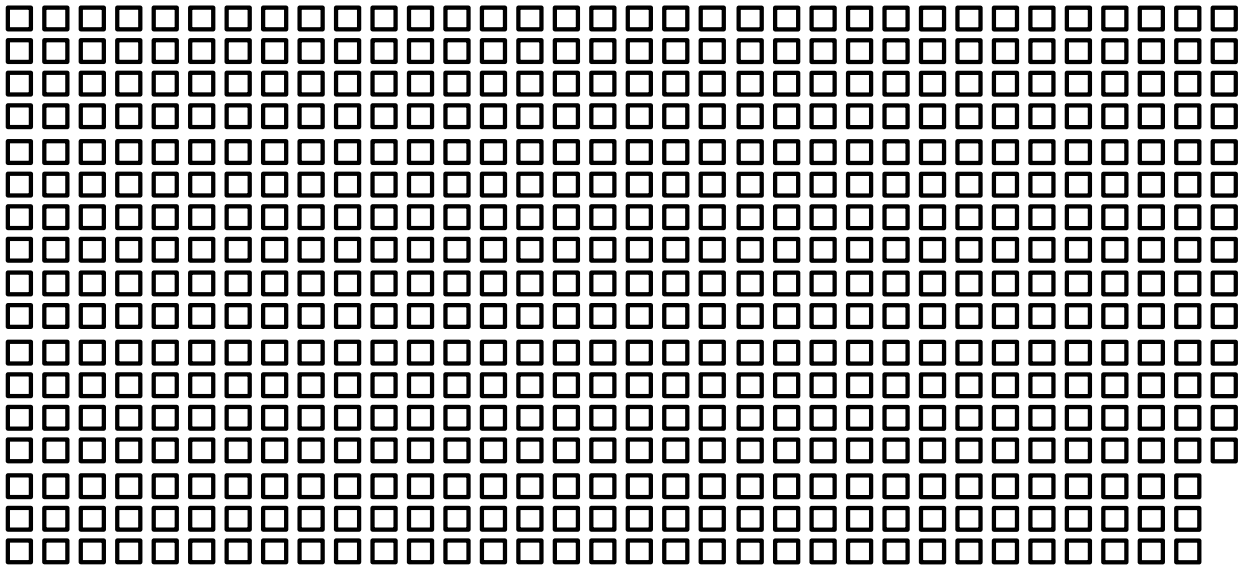
Annotating variants with SnpEff

- Annotations using reference genomes
- Calculate effects:
  - Coding (e.g. Syn, Non-Syn, Stop gained, Splice)
  - Non-coding (e.g. TFBS)
- Basic prioritizations (putative impact): {HIGH, MODERATE, LOW, MODIFIER}
- And many other things…

Pablo Cingolani

McGill

Somatic Mutations in 100 kidney tumours

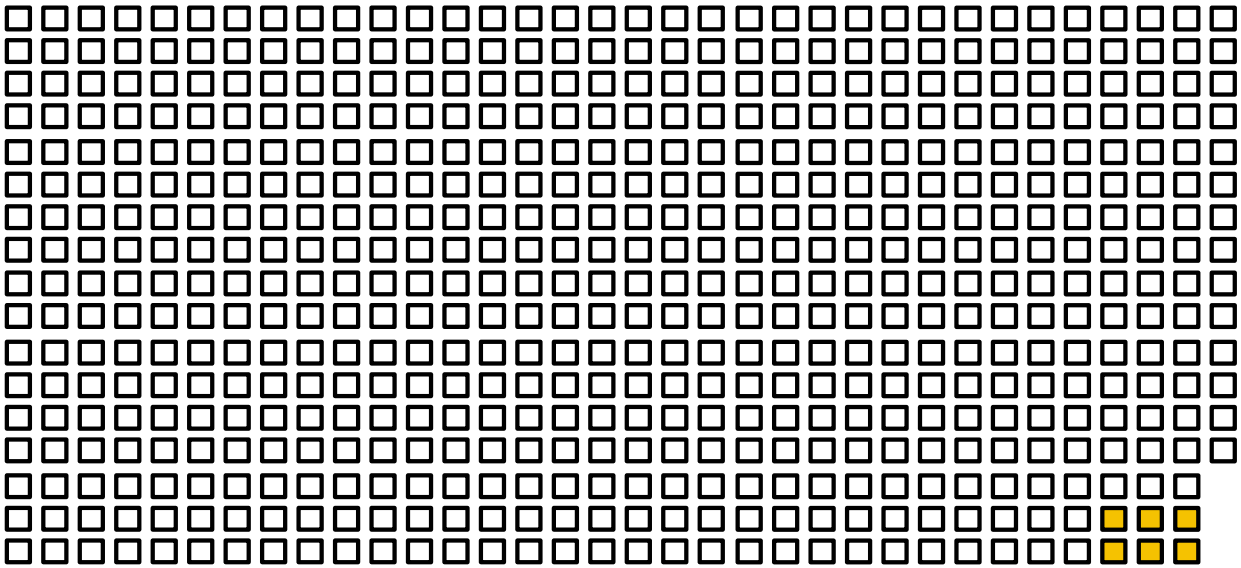


□ 1000 mutations (Total 575693)

Scelo G et al. *Nat Commun* 2014

McGill

Somatic Mutations in 100 kidney tumours



1000 mutations (Total 575693)
1000 coding mutations (Total 6172)

Scelo G et al. *Nat Commun* 2014

McGill

Conclusions

- NGS data analysis shouldn’t be a black box
- Understanding the various steps is important because choices of threshold, etc. will have an impact on data interpretation
- Without knowing all software in details it’s important to know strengths and weaknesses of each

GenAPpipes

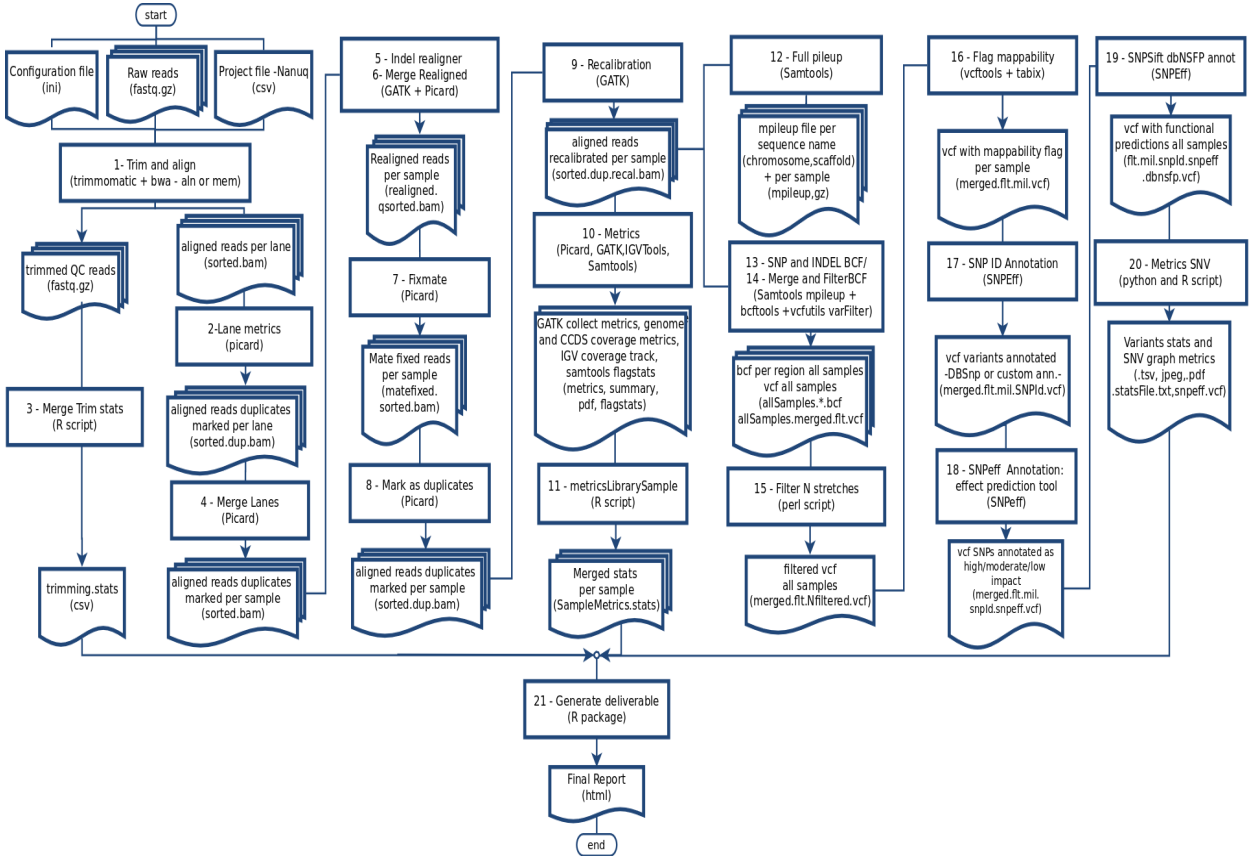
- **RNA-Seq** Spliced alignment, QC, differential analysis, isoform analysis, …
- **ChIP-Seq** Narrow/Wide peaks, Homer, GoSeq, other annotations, …
- **RNA-Seq Denovo**, differential analysis, QC, transcript annotations, …
- **DNA-Seq** Alignment, Realignment, MarkDup, Recalibration, SNV, CNV, SV, …
- **Pacbio Denovo**, bacteria and genomes up to ~50Mb, annotations (in-progress)

- All pipelines are optimised for the hardware and schedulers at the different cluster sites (the configuration is adjustable)
- All pipelines include an HTML report with the references, explanations and details on the sequencing and analysis

GenAPIPES

McGill

DNAseq overview



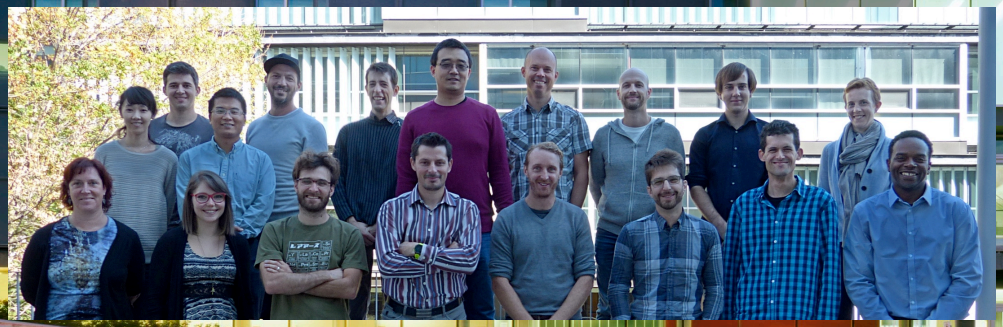
McGill

Acknowledgements

Lab
LeeAnn Ramsay
Jean Monlong
Simon Girard
Toby Hocking
Patricia Goerner-Potvin
Joe Sue
David Venuto

Analysis team
Mathieu Bourgey
Gary Lévesque
Robert Eveleigh
Francois Lefebvre
Johanna Sandoval
Pascale Marquis
Edouard Henrion

IHEC Data Portal and GenAP
David Bujold
Catherine Côté
Bryan Caron
Kuang Chung Chen
Simon Nderitu
ME Rousseau
Pierre-Étienne Jacques (UdeS)
David Morais (UdeS)
Carol Gauthier (UdeS)
Alain Veilleux (UdeS)
Maxime Levesque (UdeS)
Michel Barrette (UdeS)



guil.bourque@mcgill.ca

McGill
canarie
compute • calcul CANADA
CIHR IRSC Canadian Institutes of Health Research / Instituts de recherche en santé du Canada
GenomeCanada
TOP GLOBAL UNIVERSITY JAPAN
KYOTO UNIVERSITY FOUNDED 1897 JAPAN
GenomeQuébec