6 Applications of 2nd-Gen Sequencing

This exposition was developed by David Weese. It is based on:

1. Rasmussen, K., Stoye, J. and Myers, E. W. (2006). *Efficient q-gram filters for finding all ϵ-matches over a given length*, J. Comp. Biol.
2. Weese, D., Emde, A., Rausch, T., Döring, A. and Reinert, K. (2009). *RazerS - Fast Read Mapping with Sensitivity Control*, Genome Res.

6.1 Second-generation sequencing technologies

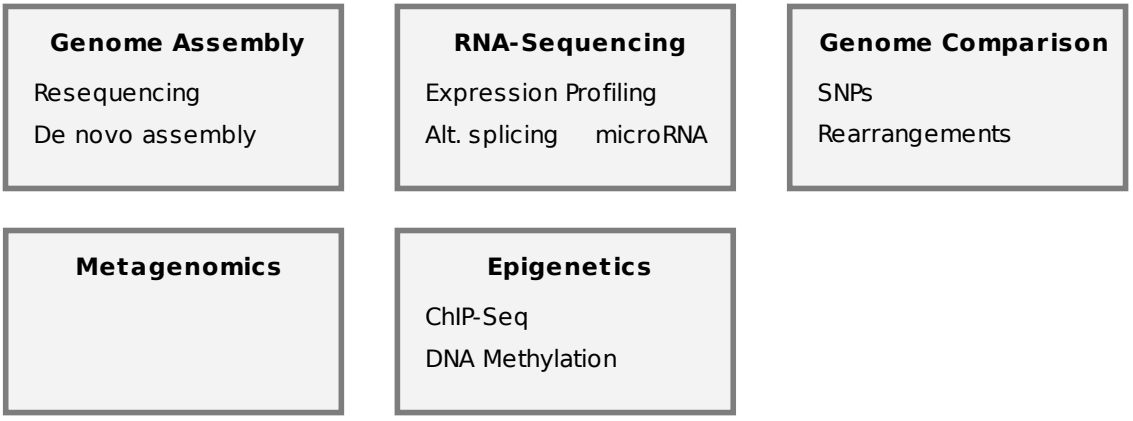
	454 FLX/Roche	Solexa/Illumina	SOLiD/ABI
Sequencing approach	pyrophosphate release	bridge amplification	ligation
Read lengths	400–500bp	36bp	35bp or 25bp (MP)
Mate pairs	yes	yes	yes
Output/Run	400–600Mbp in 10h	$>$ 1.5Gbp in 2.5d	3–4Gbp in 6d
Accuracy depends on	homopolymer length ($>$ 6 problematic)	nucleotide position in the read	nucleotide position in the read

GS FLX Titanium Series

Genome Analyzer 2

SOLiD System 2.0 Analyzer

6.2 Second-generation sequencing applications



6.3 Motivation

Fundamental to almost all of these applications is the following problem:

Problem 1 (Read Mapping Problem). *Given a set of read sequences $\mathcal{R}$, a reference sequence G, and a distance $k \in \mathbb{N}$. Find all pairs (r, g) with $r \in \mathcal{R}$, g is substring of G and $dist(r, g) \leq k$.*

Common distance measures are Hamming distance or edit distance.
The pairs (r, g) are called **matches** of r.

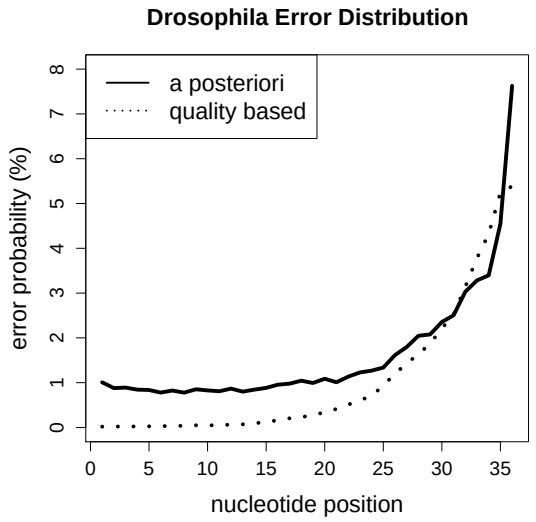
(Some of the) Existing Read Mapping Tools:

	Eland	**Maq**	**Soap**	**Seqmap**	**Zoom**	**Shrimp**	**RazerS**
filtering technique	two-seed pigeonhole	two-seed pigeonhole	two-seed pigeonhole	two-seed pigeonhole	multiple gapped seeds	q-gram counting	q-gram counting
distance measure in filtering step	Hamming	Hamming	Hamming	both	Hamming	both	both
distance measure in mapping step	Hamming	Hamming (Smith-Wa. for second mate)	Hamming (optionally with one gap)	Hamming or edit with at most 5 errors	Hamming or edit with at most one gap	Smith-Waterman	either edit or Hamming
supported read length	≤ 32	≤ 127	≤ 60	arbitrary	≤ 63	arbitrary	arbitrary
sensitivity	full sensitivity only for up to 2 errors	full sensitivity	depends on setting, no switch to guarantee full sensitivity	full sensitivity	switch to guarantee full sensitivity	no help for parameter choice, default will be lossy for most settings	arbitrarily adjustable
can output all (suboptimal) hits	no	no	no	yes	yes	yes	yes

Observation 2 (Sharpness)**.** *The q-gram Lemma is sharp. The worst case occurs if the k errors are* equidistantly distributed.

The error probability increases with the nucleotide position for Illumina, SOLiD, and Sanger sequencing.

Observation 3. *The worst case occurrence probability is* very small.



Drosophila melanogaster reads (NCBI short read archive, SRR001815, Illumina tech.)

Why not increasing t or q to increase filtration specificity and reduce runtime?
How many matches would be lost?

Definition 4. Sensitivity is the probability that a true match is classified as potential match:

$$P(\#\text{matching } q\text{-grams} \geq t \mid \#\text{errors} \leq k)$$

Loss Rate = 1 − Sensitivity

6.4 How to calculate the sensitivity

Let $p_i^{\mathtt{R}}$ be the probability of a sequencing error at nucleotide position i.

We could enumerate all configurations of $e = 0, \dots, k$ errors and sum up the occurrence probs of those with $\geq t$ matching q-grams.

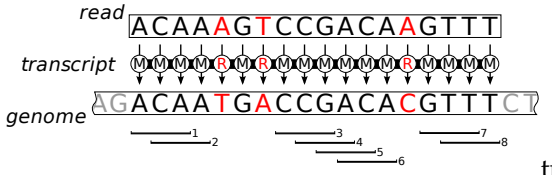
```
genome   ACAAAGTCCGACAAGTTT
         |||| | ||||||  ||||
read     ACAATGACCGACACGTTT      match with 3 replacements
```

The occurrence probability would be $p_1^{\mathtt{m}} p_2^{\mathtt{m}} p_3^{\mathtt{m}} p_4^{\mathtt{m}} p_5^{\mathtt{R}} p_6^{\mathtt{m}} p_7^{\mathtt{R}} p_8^{\mathtt{m}} \dots p_{13}^{\mathtt{m}} p_{14}^{\mathtt{R}} p_{15}^{\mathtt{m}} p_{16}^{\mathtt{m}} p_{17}^{\mathtt{m}} p_{18}^{\mathtt{m}}$, with $p_i^{\mathtt{m}} = 1 - p_i^{\mathtt{R}}$.

The enumeration would take $\Omega\left(\left(\frac{n}{k}\right)^k\right)$ time. Not feasible for $n = 200$ and $k = 20$ errors.

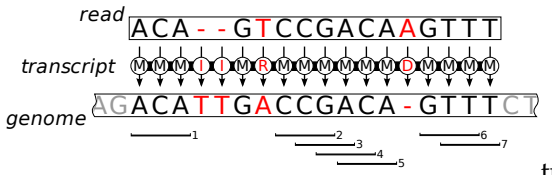
6.5 Definitions

- Considering mismatches only (Hamming distance)



transcript $T \in \Delta^*$, $\Delta = \{\mathtt{M}, \mathtt{R}\}$, $|T| = |read|$

- Considering mismatches and indels (edit distance)



transcript $T \in \Delta^*$, $\Delta = \{\mathtt{M}, \mathtt{R}, \mathtt{I}, \mathtt{D}\}$,
$|read| \leq |T| \leq |read| + \#errors$

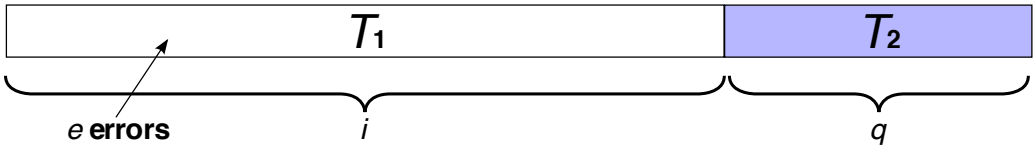
- Substrings $\mathtt{M}^q$ are called *q-matches*
- A q-match corresponds to a common q-gram

6.6 DP Approach (Hamming)

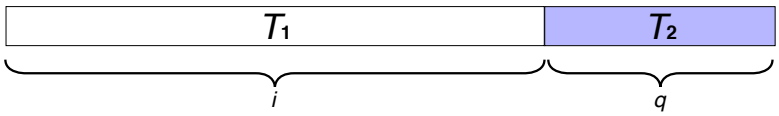
- Much more efficient than the full enumeration
- Recursive enumeration of all error configurations explicitly storing only the last q positions
- DP-Matrix R with $R(i, e, t, T_2)$

i	=	1st transcript length	$0, \dots, \lvert read\rvert - q$
e	=	number of errors	$0, \dots, k$
t	=	threshold	$0, \dots, r - q + 1$
T_2	=	2nd transcript of length q	$T_2 \in \Delta^q$ where $\Delta = \{\mathtt{M}, \mathtt{R}\}$

- Contains the sum of occurrence probabilities of transcripts T_1 s.t. T_1 contains e letters $\mathtt{R}$, $T_1 T_2$ contains $\geq t$ substrings $\mathtt{M}^q$

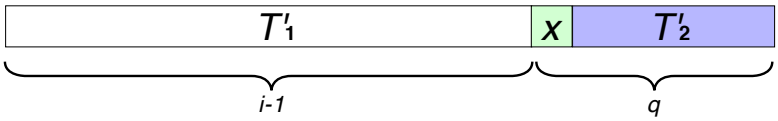


The sum of occ. probs. of all transcripts T_1 ...



T_1 contains e errors
T_1T_2 contains $\geq t$ substrings $\mathtt{M}^q$

... can be computed recursively from



with $x \in \{\mathtt{M}, \mathtt{R}\}$
$T'_1 = T_1[1..i-1]$
$T'_2 = T_2[1..q-1]$

$$T'_1 \text{ contains } \begin{cases} e, & \text{if } x = \mathtt{M} \\ e-1, & \text{if } x = \mathtt{R} \end{cases} \text{ errors}$$

$$T'_1 x T'_2 \text{ contains } \geq \begin{cases} t-1, & \text{if } T_2 = \mathtt{M}^q \\ t, & \text{else} \end{cases} \text{ substrings } \mathtt{M}^q$$

Lemma 5 (Hamming distance, ungapped)**.**

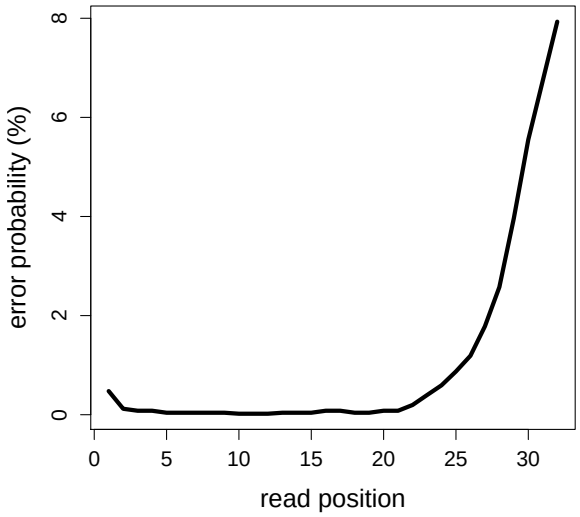
$$R(0,e,t,T_2) = \begin{cases} 1, & \text{if } e = 0,\ t \leq \delta(T_2) \\ 0, & \text{else} \end{cases}$$

$$\begin{aligned} R(i,e,t,T_2) = {} & p_i^{\mathtt{M}} \cdot R(i-1, e, t-\delta(T_2), \mathtt{M}T_2[1..q-1]) \\ + {} & p_i^{\mathtt{R}} \cdot R(i-1, e-1, t-\delta(T_2), \mathtt{R}T_2[1..q-1]) \end{aligned}$$

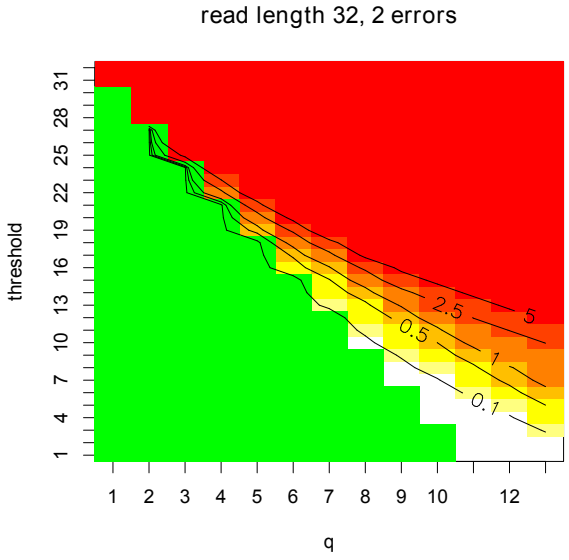
$$\delta(T) := \begin{cases} 1, & \text{if } T = \mathtt{M}^q \\ 0, & \text{else} \end{cases}$$

We devised a DP algorithm that calculates the sensitivities for all $e = 0,\dots,k$ and $t = 1,\dots,t_{\max}$ in $O(n \cdot k \cdot t_{\max} \cdot 2^q)$.

The recursion can be extended to gapped shapes and edit distance.

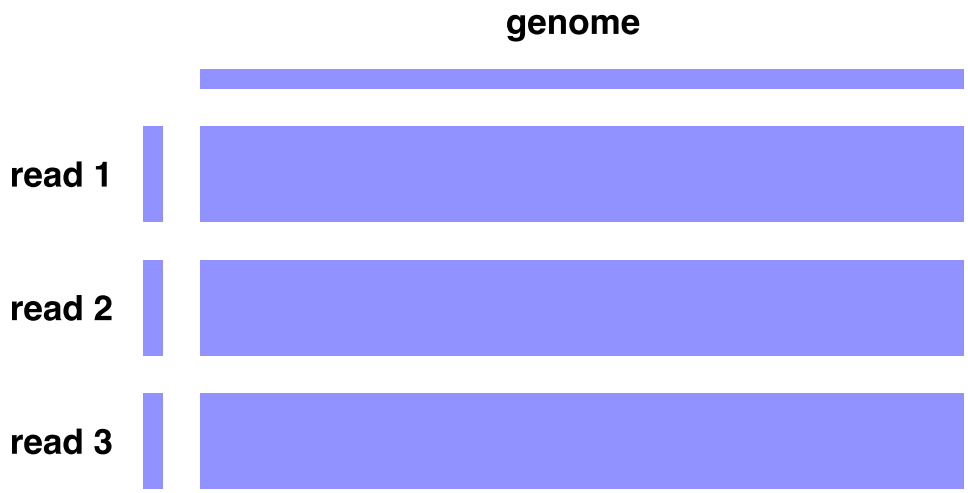


Example error distribution

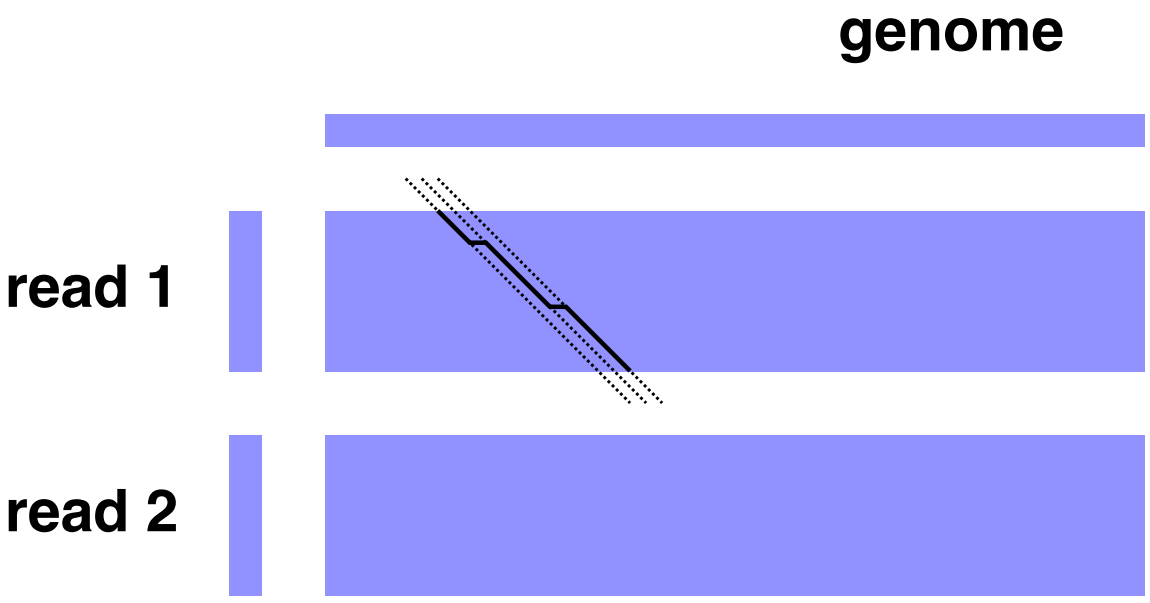


Different loss rates of matches with 2 replacements

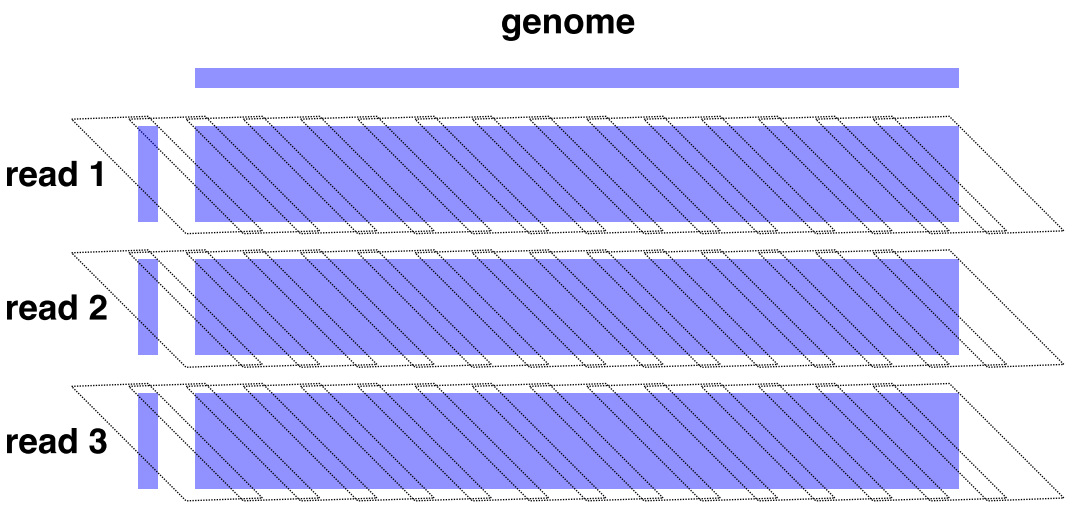
6.7 Filtration



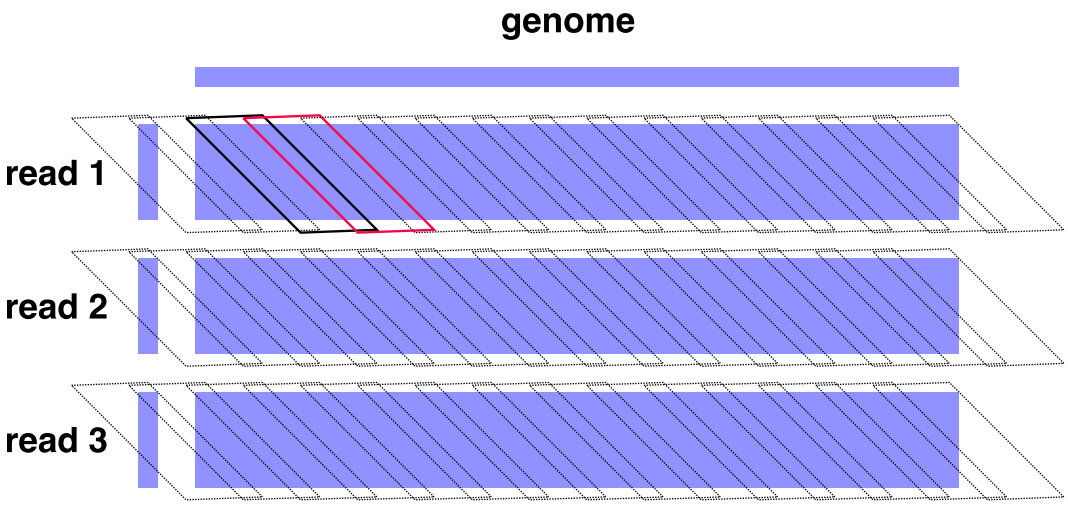
Consider dotplots between genome and reads.



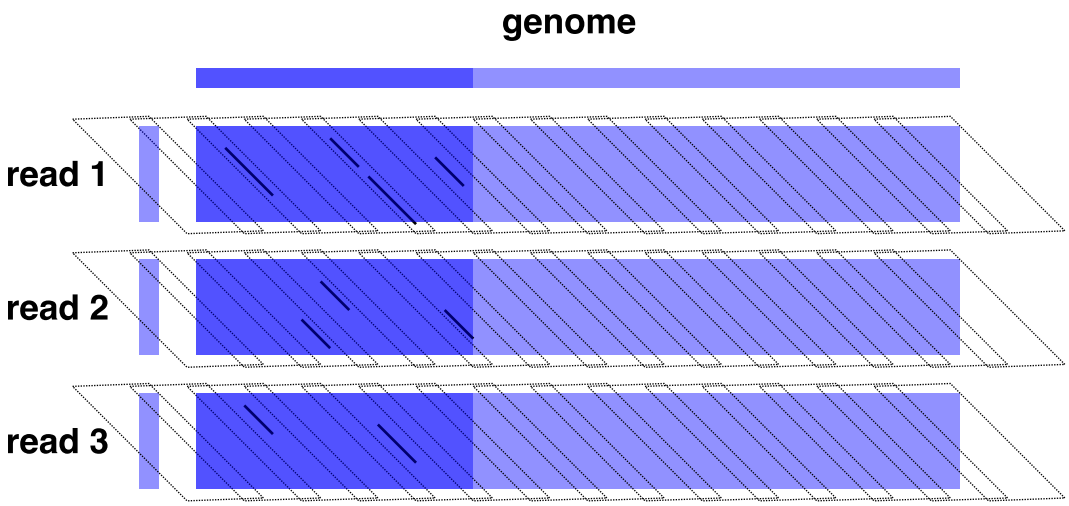
A match with k indels is covered by at most $k + 1$ consecutive diagonals.



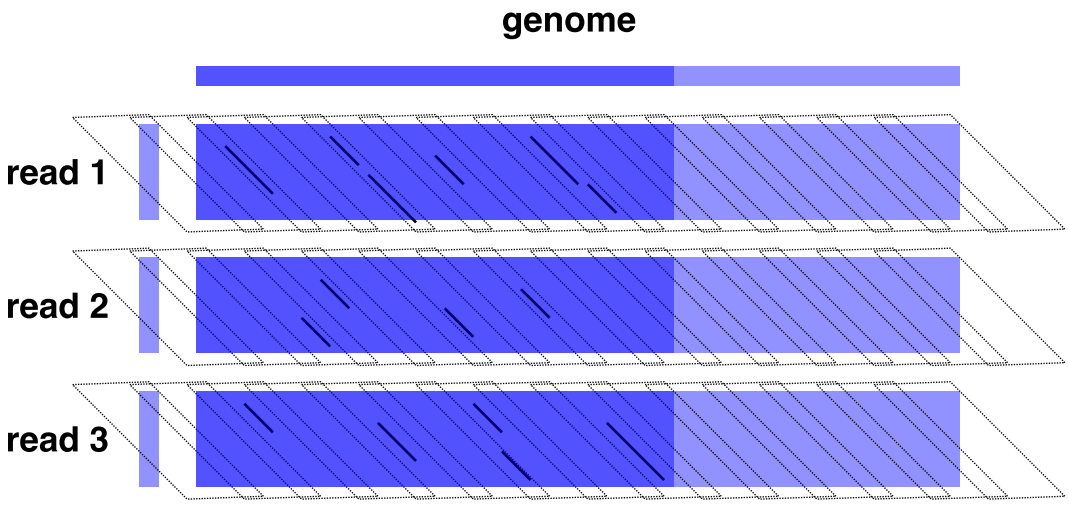
Cover the dot plots with parallelograms of w diagonals with $w \geq k + 1$ ...



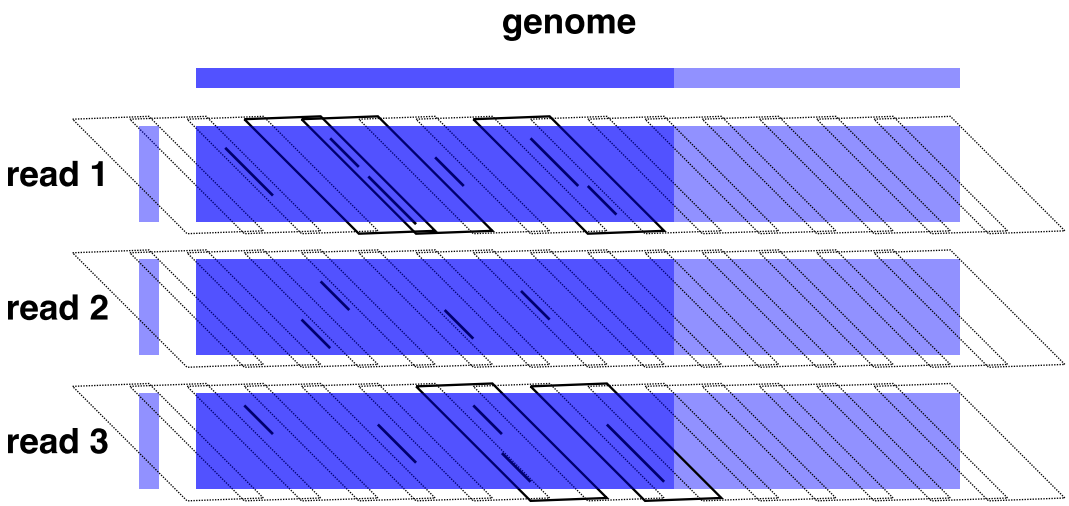
... and an overlap of k diagonals.
Every possible sequence of $k + 1$ diagonals resides in a parallelogram.



Search common q-grams while scanning the genome from left ...



... to right.



Potential matches are contained in parallelograms with $\geq t$ common q-grams.

Optimizations as suggested in [RasStoMye05]:

- Associate every parallelogram with a counter
- Reuse counters after $(|r| - w - q)$ q-gram sliding steps
- Choose parallelogram width, s.t. they begin at multiples of a power of 2 → Fast bit-shift reveals counter number of a diagonal

Differences compared to Swift:

- not local, but semi-global alignment
- parallelograms are not opened or closed, they are verified as a whole

6.8 Verification

- Edit Distance
  - Parallelograms are verified with bit-vector algorithm by [Myers99]
  - It exploits hardware parallelism of bit-operations:
    A 64-bit CPU calculates 64 DP cells in 14 arithmetic/logic operations
  - Returns the end position of a true match in the genome
  - Can be modified to also return the beginning

- Hamming Distance
  - Scan each diagonal until $k + 1$ mismatches occur
  - Every diagonal with $\leq k$ mismatches is a true match

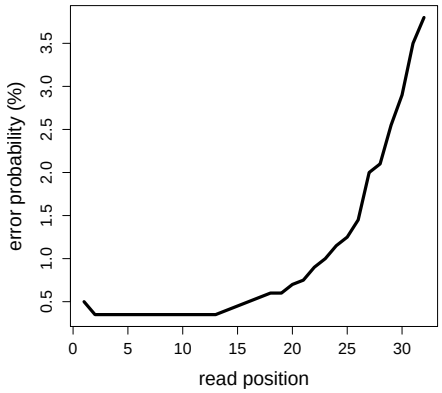
6.9 Filtration Parameter Choice

Automatically choose filtering parameters (shape Q and threshold t) to ...

- achieve a certain sensitivity level
- minimize the running time of the mapping procedure

Therefore precompute the loss rates for ...

- read lengths from 24 to 100
- error rates up to 10%
- a typical Illumina error profile [Dohm08]



Parameters for larger reads are extrapolated from precalculations with the same error rate.

Parameter tables can be recomputed with user-specific error distribution from:

- **Quality based probabilities:** Transform the average base call quality value for each position into a probability value.
- **A posteriori probabilities:** Map a small subset of reads and determine the position dependent error frequency.

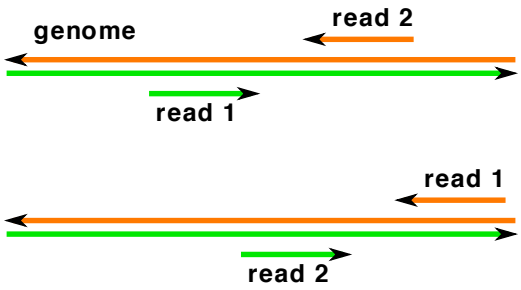
For instance, parameters of 50bp reads can be recalculated within 10min using the DP algorithm.

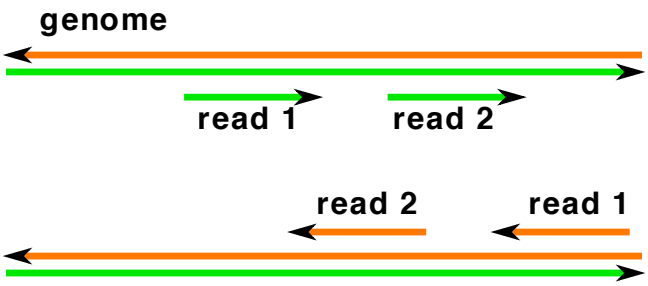
6.10 Paired-End Read Mapping

Given a library size μ and a tolerated deviation δ, we want to find all paired-end matches with

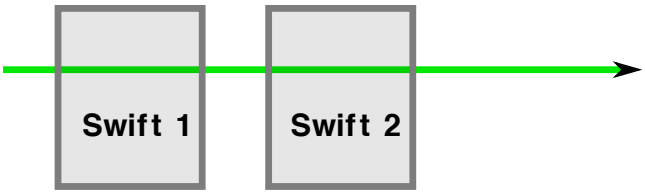
- an insertion size between $\mu - \delta$ and $\mu + \delta$,
- each mate matches with up to k errors.

Paired-end reads are sequenced from different strands and "look at each other".
There are two symmetries:

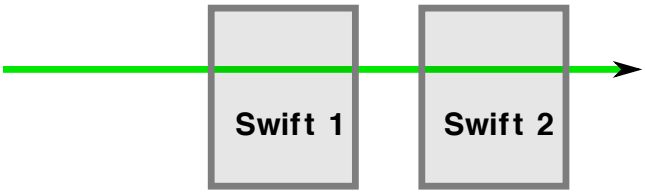




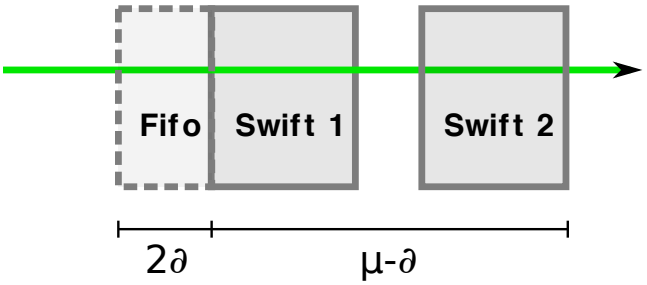
Search read 1 and the reverse-complemented read 2 on the **same** strand.



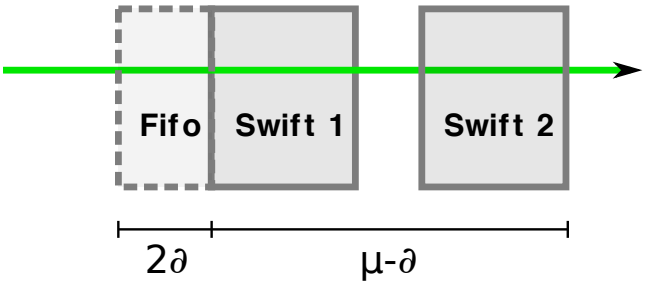
Scan one strand with two Swift filters from left . . .



. . . to right.



Keep their distance and record the potential matches of the trailing filter in a fifo.



Each true paired-end match is recognized as a Swift 2 potential match and a potential mate match in the fifo.

Implementation:

- As optimization use a last-seen-at table for Swift 1 potential matches.
- Output a true match with minimal errors and a minimal library deviation.

6.11 Results

The following datasets were used:

Read sets[1] from the NCBI short read archive:

- 10,760,364 × 36bp reads of Drosophila melanogaster (SRR001815)
- 7,894,743 × 2 × 76bp reads of a human HapMap individual (SRR006387) trimmed to 63bp (Zoom's limit)

Reference genomes:

- Drosophila melanogaster genome from FlyBase, Release 5.9
- Human genome from NCBI, Build 36.3

Verification of Expected Sensitivity:

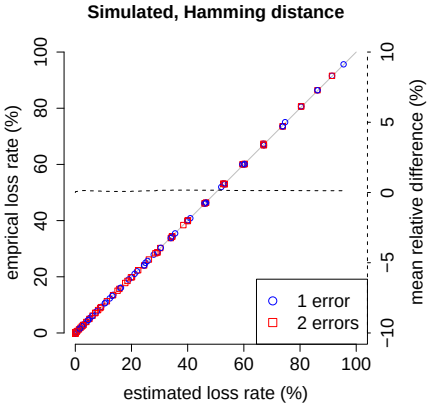
[1] In both sets, the Illumina technology was used.

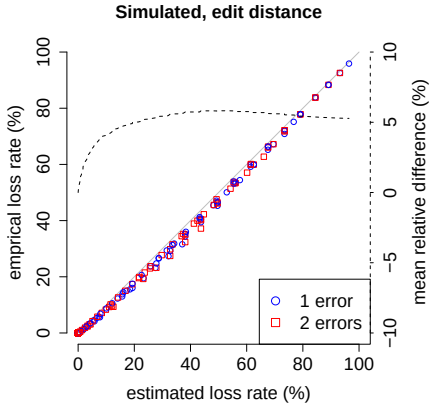
1. Simulated reads
   - Simulate reads using a positional error model [Dohm08]
   - Group them according number of errors
   - **Empirical sensitivity** is the proportion that could be mapped back to origin

2. Real data
   - Map Drosophila reads lossless and keep those, that map uniquely
   - Group them according number of errors
   - Determine positional error model
   - **Empirical sensitivity** is the proportion that could be mapped back to origin

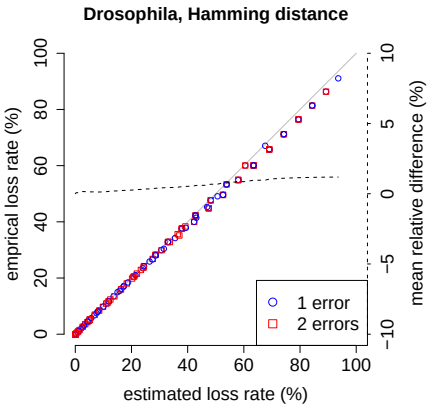
For different filtration settings ($q = 8...14$, $t = 1...20$) we calculated the **estimated loss rate** and compared it with the **empirical loss rate**.

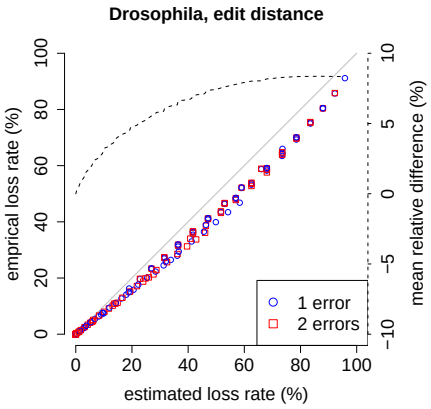
1. Simulated reads





2. Real data





Short Read Experiments:

1. *Drosophila, Hamming distance 2*
   - Map Drosophila reads onto the Drosophila genome with ≤ 2 mismatches
   - Shrimp uses Smith-Waterman verifier $\rightarrow$ we adapted the scores (mismatch $= 0$, match $= 1$, gap penalties $= -1000$, score threshold $= 34$).

2. *Drosophila, edit distance 2.*
   - As in experiment (1) allowing also indels
   - Shrimp computes local alignments $\rightarrow$ no scoring scheme for semi-global edit distance alignments
   - We emulated edit distance with (mismatch $= -1$, match $= 1$, gap penalties $= -1$, score threshold $= 32$)

3. *Human, Hamming distance 5.*
   - Map HapMap reads onto the human genome with ≤ 5 mismatches
   - Adapted Shrimp scores as in experiment (1)

experiment			RazerS100	RazerS99	Zoom	Shrimp	SeqMap	Soap	Maq
(1)	1M	time (min)	2.13	1.63	**1.47**	15.3	6.70	9.27	4.10
		space (GB)	1.31	1.30	0.72	0.68	6.56	0.67	0.60
		mapped reads	505,506	503,595	505,506	505,084	505,059	506,476	503,999$^{(605K)}$
	all	time (min)	10.6	**5.55**	7.80	145	12.8	116	9.68
		space (GB)	4.10	3.92	3.77	5.80	11.1	0.67	5.36
		mapped reads	5,353,287	5,335,554	5,353,287	5,349,007	5,348,776	5,414,337	5,338,676$^{(6.5M)}$
(2)	1M	time (min)	12.3	**5.92**	32.7	13.6	15.5		
		space (GB)	0.48	0.53	0.72	0.68	8.38	-	-
		mapped reads	512,477	511,695	512,139	515,080	512,477		
	all	time (min)	163	**68.45**	267	146	abort		
		space (GB)	4.58	4.59	3.77	5.90		-	-
		mapped reads	5,431,142	5,424,088	5,427,589	5,486,467			
(3)	1M	time (h)	3.14	**0.40**	26.1	10.7	48.8	3.88	2.43
		space (GB)	1.14	1.86	1.27	6.10	8.10	6.20	0.70
		mapped reads	352,725	351,767	352,617	352,742	349,721	354,020	323,893$^{(362K)}$
	all	time (h)	25.4	**1.95**	45.3	> 3 d	abort	33.8	5.74
		space (GB)	5.60	6.13	2.89			6.2	4.38
		mapped reads	3,102,320	3,095,435	3,091,063			3,133,920	2,817,561$^{(3.0M)}$

Paired-end read mapping onto unmasked human chromosome 21

- $2 \times 1{,}000{,}000$ and $2 \times 7{,}894{,}743$ reads of length 63
- up to 5 mismatches
- full and 99% sensitivity

			RazerS100	RazerS99	Zoom	Maq
Paired-end	1M	time (min)	36.1	6.45	**3.38**	11.5
		space (GB)	1.28	3.13	2.59	0.85
		mapped pairs	26,923	26,828	14,018	19,025$^{(28.5K)}$
	all	time (min)	71.4	47.5	**22.2**	72.9
		space (GB)	10.8	12.5	20.5	4.72
		mapped pairs	241,308	240,385	129,704	167,015$^{(238K)}$

Read mapping onto unmasked human chromosome 21

- 500,000 simulated 125bp and 250bp reads
- up to 8% errors, full and 99% sensitivity

			RazerS100	RazerS99	Shrimp100	Shrimp99	Maq
Hamming	125bp	time (min)	8.53	**4.15**	9.61 h	60.9	4.54
		space (GB)	0.80	1.54	1.93	4.48	0.38
		mapped	500,000	499,991	500,000	499,991	405,377
	250bp	time (min)	14.7	**6.65**	32.9 h	160	
		space (GB)	1.26	2.00	1.46	2.61	-
		mapped	500,000	500,000	500,000	500,000	
edit	125bp	time (min)	65.0	**23.7**	9.44 h	61.6	
		space (GB)	0.74	1.71	0.84	4.50	-
		mapped	500,000	500,000	500,000	500,000	
	250bp	time (min)	55.8	**39.6**	14.7 h	28.5 h	
		space (GB)	1.21	2.18	1.38	5.45	-
		mapped	500,000	499,940	500,000	499,940	

6.12 Second-generation sequencing applications

Genome Assembly
Resequencing
De novo assembly

RNA-Sequencing
Expression Profiling
Alt. splicing microRNA

Genome Comparison
SNPs
Rearrangements

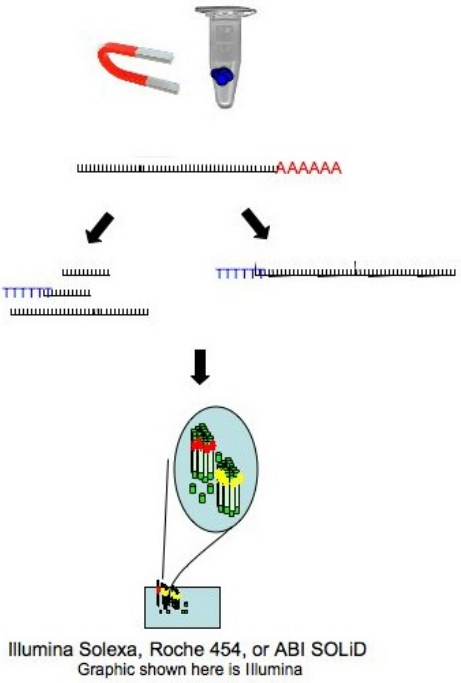
Metagenomics

Epigenetics
ChIP-Seq
DNA Methylation

6.13 RNA-Sequencing

How RNA-Seq works:

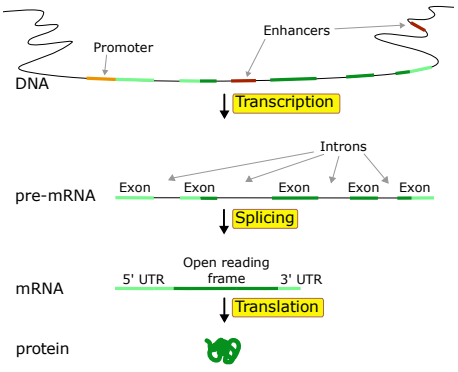
- RNA isolatation
- Reverse transcription to cDNA
- Fragmentation
- (Size selection)
- Sequencing

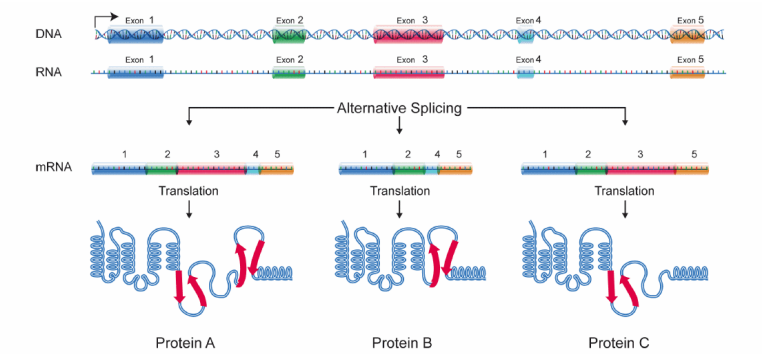


RNA-Seq applications:

- **Expression profiling:** Quantify gene expression levels
- **Alternative splicing:** Which mRNAs are generated from the same gene?
- **microRNA:** Where is the genomic source, which genes are regulated?

6.14 RNA-Seq - Alternative Splicing





Two approaches two determine splice variants:

1. Cut the genome at known splice sites and map mRNA reads onto combinations of merged genome fragments
2. Map as many mRNA reads as possible onto the genome and use coverage and known introns to detect new splice sites. Proceed as above.

Second Approach[a]:

- Map reads
- Assemble uniquely mapped reads
- Generate possible splices
- Try to map the non-uniquely mapped reads onto splices

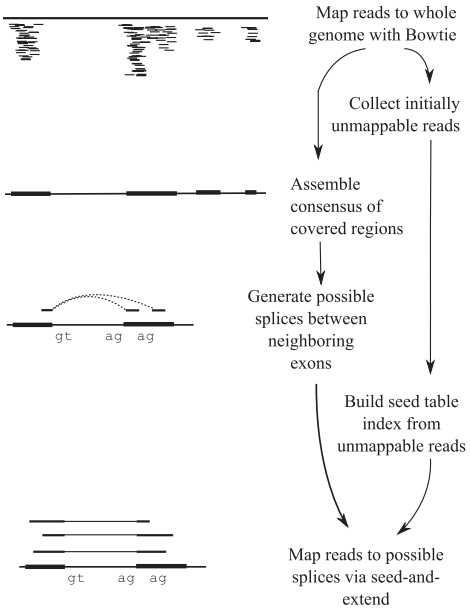
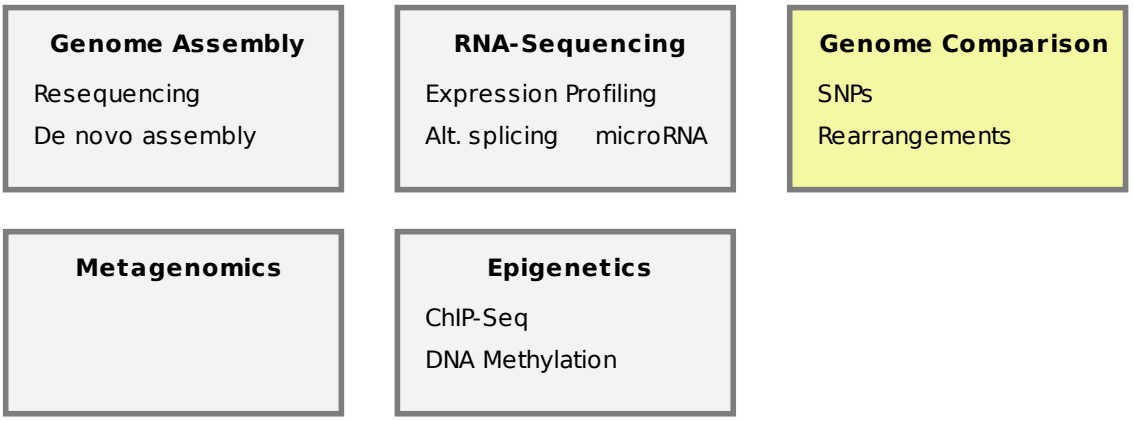


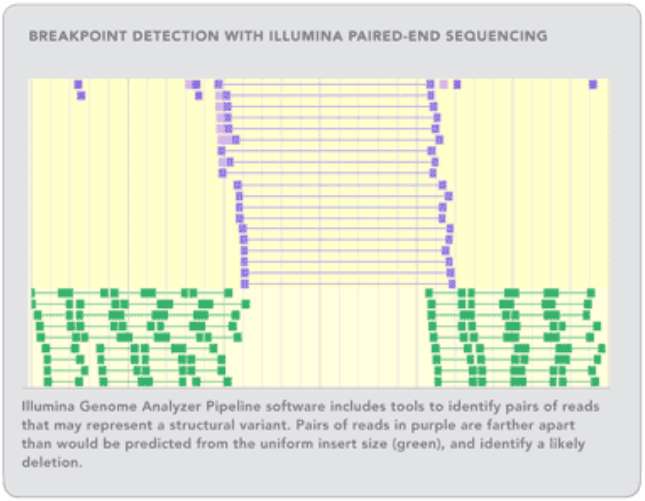
Fig. 1. The TopHat pipeline. RNA-Seq reads are mapped against the whole reference genome, and those reads that do not map are set aside. An initial consensus of mapped regions is computed by Maq. Sequences flanking potential donor/acceptor splice sites within neighboring regions are joined to form potential splice junctions. The IUM reads are indexed and aligned to these splice junction sequences.

[a] Trapnell C, Pachter L, Salzberg SL. (2009) *TopHat: discovering splice junctions with RNA-Seq*, Bioinformatics

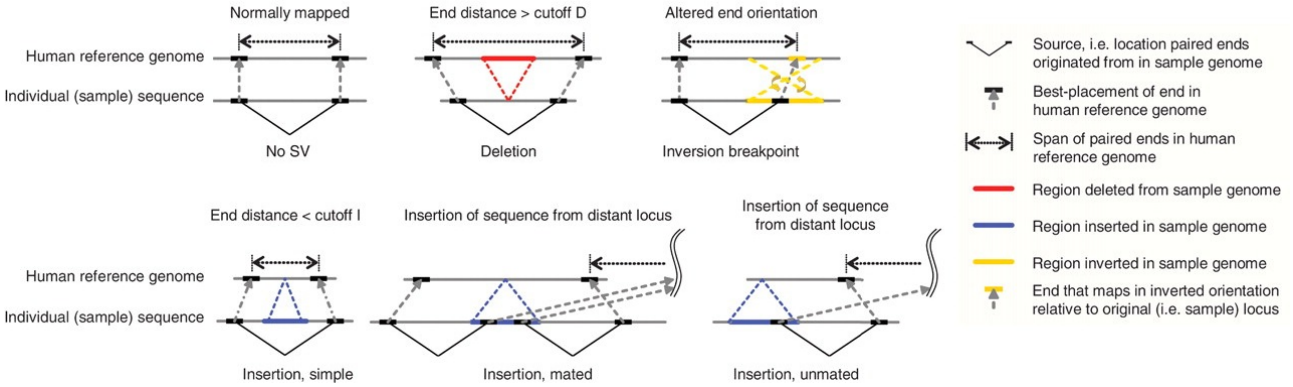


6.15 Genome Comparison

- Sequence paired-end reads of an unknown genome (sample)
- Map them onto a known reference genome (target)
- Search for small mutations (SNPs) or large structural variations (rearrangements) between them



A deletion in the sample induces pairs of reads to be farther apart than predicted.

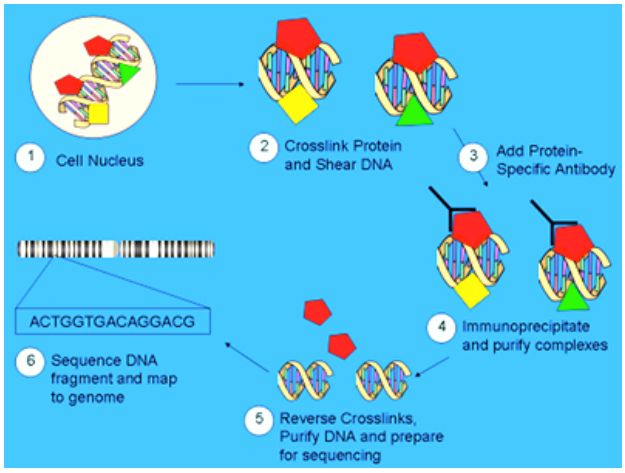


Inversions, deletions, translocations can also be detected.[23]

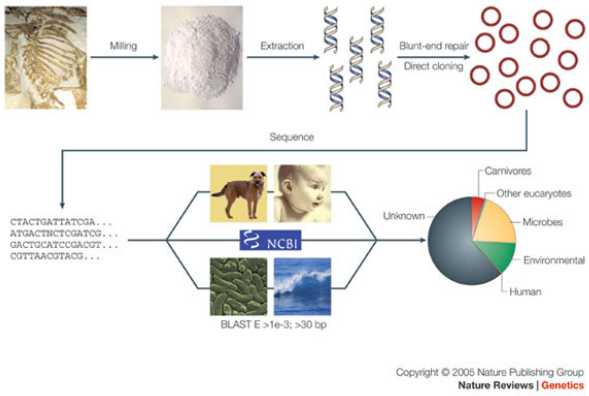
[2]Korbel JO, Urban AE, Affourtit JP, et al. (2007) *Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome*, Science

[3]Bashir A, Volik S, Collins C, Bafna V, Raphael BJ. (2008) *Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer*, PLoS computational biology

6.16 Other applications



ChIP-Sequencing[4]



Metagenomics[5]

6 Selected Publications

1. Pop, M. and Salzberg, S. L. (2008) *Bioinformatics challenges of new sequencing technology*, Trends in Genetics.
2. Mardis, E.R. (2008) *The impact of next-generation sequencing technology on genetics*, Trends in Genetics

[5]Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) *High-resolution profiling of histone methylations in the human genome*, Cell

[5]Poinar HN, Schuster SC, et al. (2006) *Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA*, Science