Mapping short paired-end reads with SSAHA2

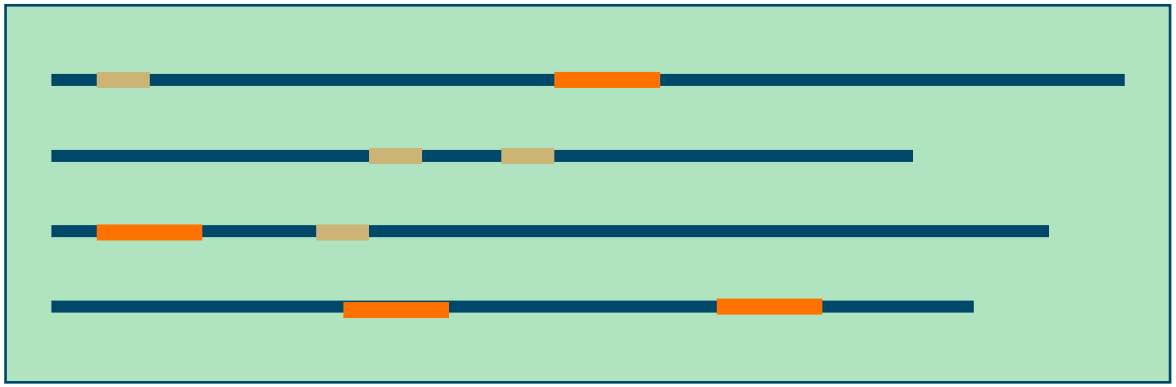
Hannes Ponstingl
hp3@sanger.ac.uk

Sequence Assembly & Analysis Group

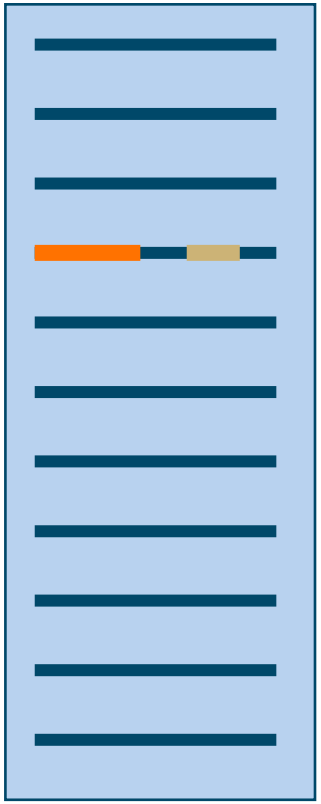
wellcome trust sanger institute

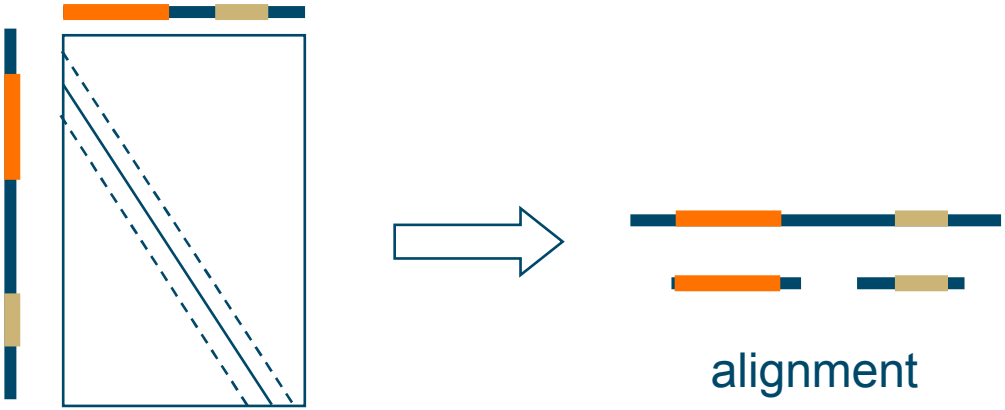
SSAHA2 mapping strategy

subject sequences (hashed k-tuples)



FASTQ file with query sequences





banded Smith-Waterman

alignment

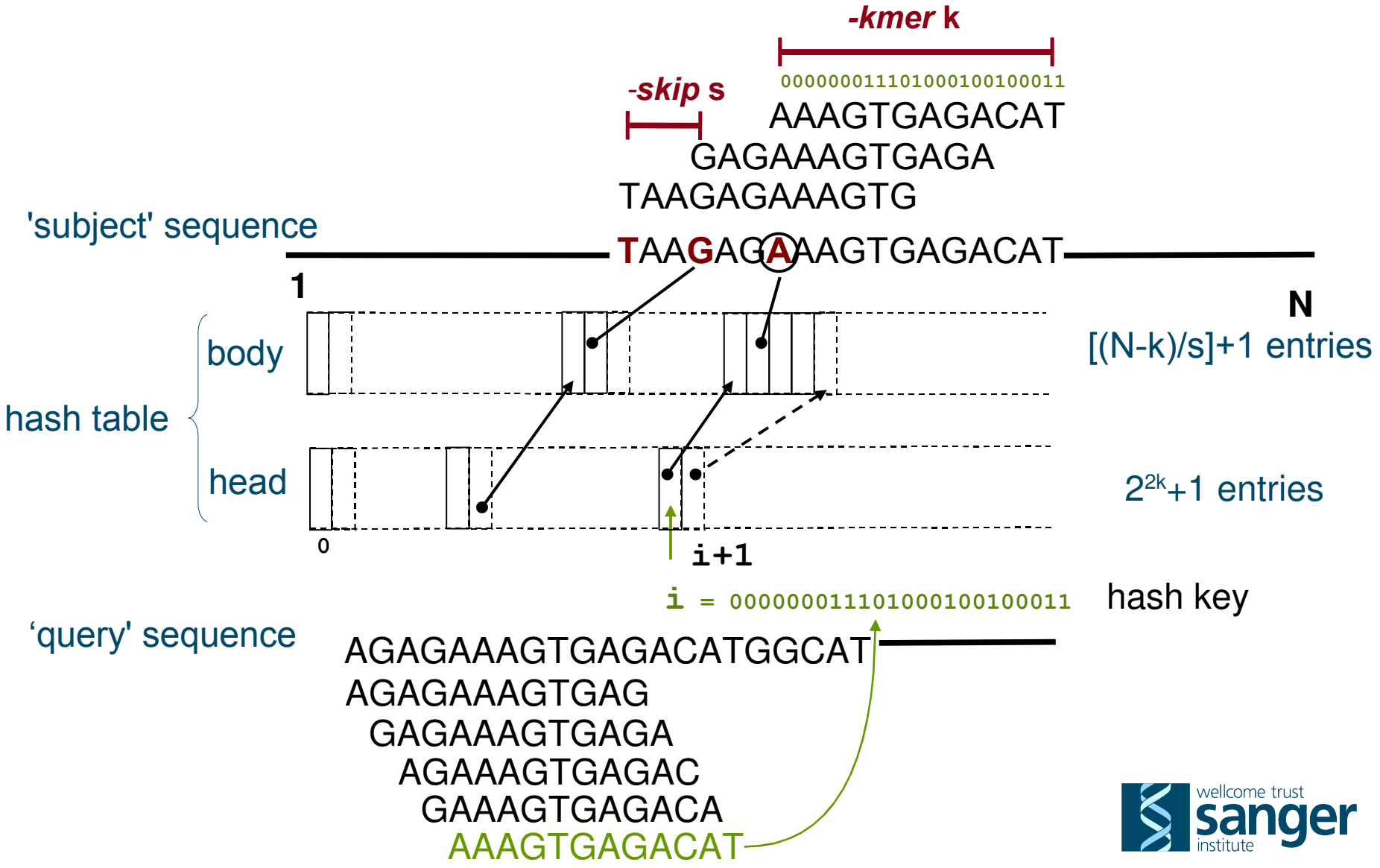
wellcome trust sanger institute

Outline

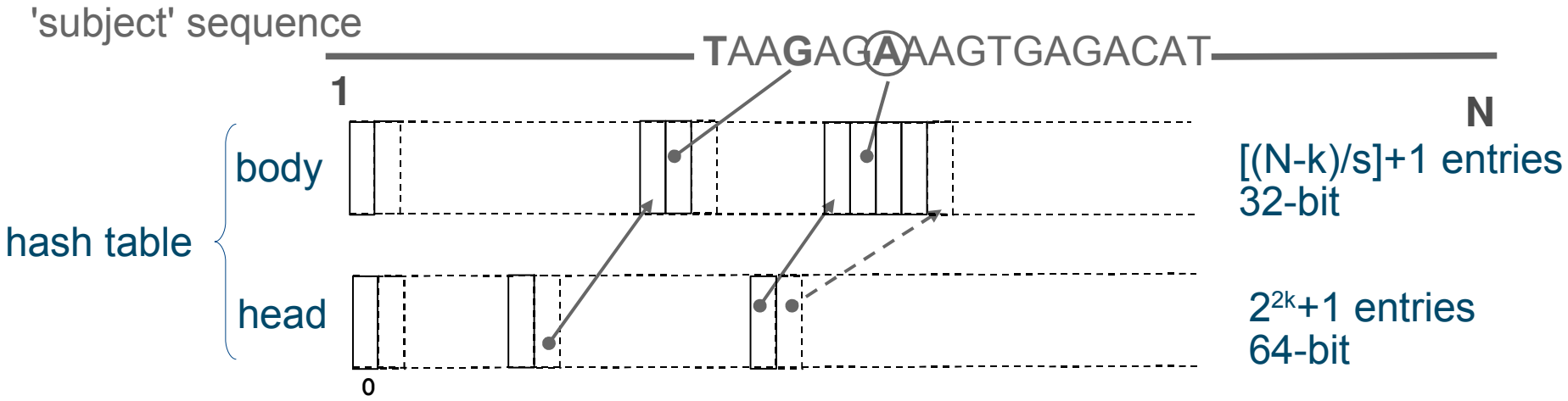
- strategy:
    1) identify potentially matching segments via short exact matches (k-tuple seeds)
    2) align those segments to query by banded Smith-Waterman
- SSAHA k-tuple hashing
- memory requirements
- how potentially matching segments are detected
- how potentially matching segments are filtered
- command line options that affect speed/sensitivity
- mapping score and base quality
- module for paired end reads
- command line examples

wellcome trust sanger institute

SSAHA k-tuple hashing



SSAHA2 memory requirements



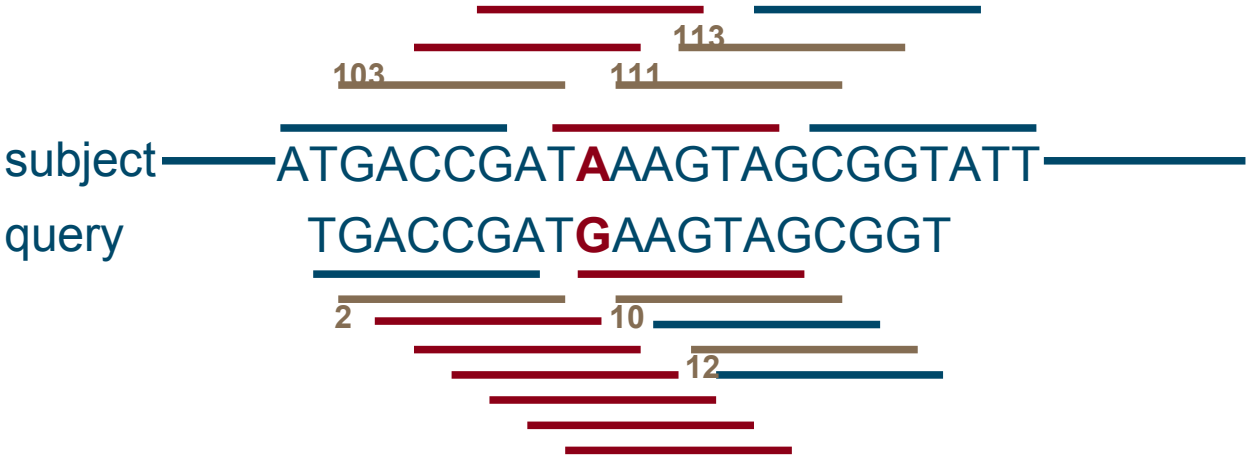
Example: human genome

ssaha2 -kmer 13 -skip 2	head: 537 MB, body: 6 GB
ssaha2 -kmer 12 -skip 1	head: 134 MB, body : 12 GB

wellcome trust sanger institute

Identify potentially matching segments (1)

> ssaha2 -kmer 7 -skip 2



k-tuple hit list (seeds):

query q	subject s	shift (s-q)
2	21	19
12	33	21
1	56	55
8	73	65
11	86	75
2	103	101
10	111	101
12	113	101
7	204	197
1	311	310
5	342	337
13	401	388
3	396	393
5	413	418
2	489	487
4	0	
6	0	
9	0	

- sort seeds by shift (max. hit list length: 4×10^6)
  → bottle-neck for large genomes in terms of speed
- join seeds of similar shift -> potentially matching segments

command line options:

-seed <n>: require a minimum number of *s* seeds for a potentially matching segment to register ($n \geq 2$)

-cut <c> : Disregard k-tuples with more than *c* hits across the genome (default: c = 10^4)

wellcome trust sanger institute

Identify potentially matching segments (2)

Implementation feature for increased speed:

- only register a potentially matching segment
  if the number of seeds is not less than *m* below the *current* maximum

- potentially matching segments are extended by *e* bases (fixed) and submitted to Smith-Waterman alignment
- option *-edge <e>* is obsolete as of version 2.0

-rtype	*m*	*e*
solexa	4	30
454	6	250
abi	8	500

wellcome trust sanger institute

Important SSAHA2 command line options

option	range	default	description
-rtype <typ>	[solexa, 454, abi]	abi	Tunes algorithm for read-types
-kmer <k>	2 < k < 16	13	k-tuple length
-skip <s>	s ≥ 1	[1,3,k]	distance between successive k-tuples in the genomic sequence
-seed <n>	n ≥ 2	[2,2,5]	Minimum number of seeds for a potentially matching segment
-cut <c>	c ≥ 2	1×10^4	Threshold in the number of hits of a k-tuple across the genome
-depth <d>	d ≥ 500	500	Submit at most the *d* segments with the highest number of seeds to Smith-Waterman

wellcome trust sanger institute

SSAHA2 mapping score

The uniqueness of a match is assessed by the difference between the two matches with the highest Smith-Waterman alignment scores S_{max} and S_{max2}:

$$\delta = (S_{max} - S_{max2}) \frac{300}{R}$$

where R is the read length. If there are multiple best matches (δ = 0), the Smith-Waterman score of the match with the lowest base quality averaged over the mismatch positions is incremented by 1 (δ' = 300/R)

The mapping score S_{map} is capped at a value of 50:

$$S_{map} = \begin{cases} \delta : \delta \leq 50 \\ 50 : \delta > 50 \end{cases}$$

i.e. $0 \leq S_{map} \leq 50$ with $S_{map} = 0$ indicating multiple matches with identical Smith-Waterman scores (ambiguous mapping).

cigar::50 MAL2_000000213299.F 1 36 + MAL2 213252 213287 + 30 M 36

wellcome trust sanger institute

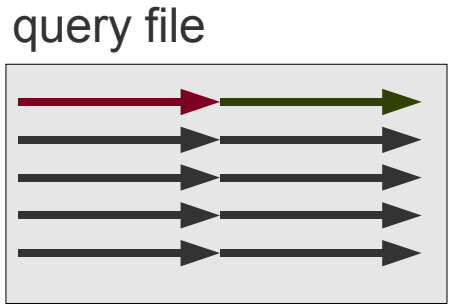
Module for paired-end reads

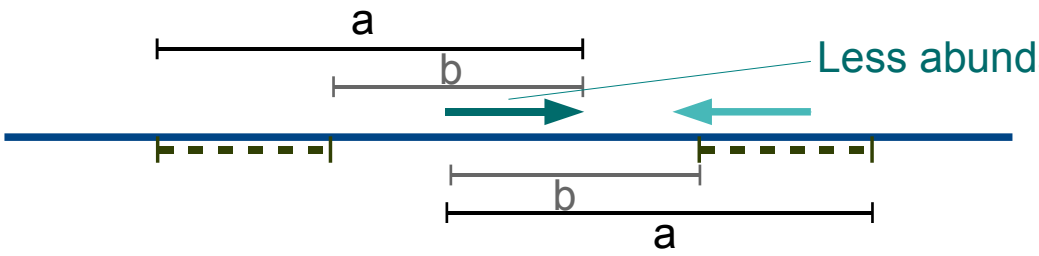
Command line option: *-pair <a,b>*

i) determine which of the mates gives rise to fewer k-tuple hits (less abundant)

ii) get matches with the best Smith-Waterman scores for that mate

iii) look in the distance band [a,b] around those positions for matches of the other, more abundant, mate



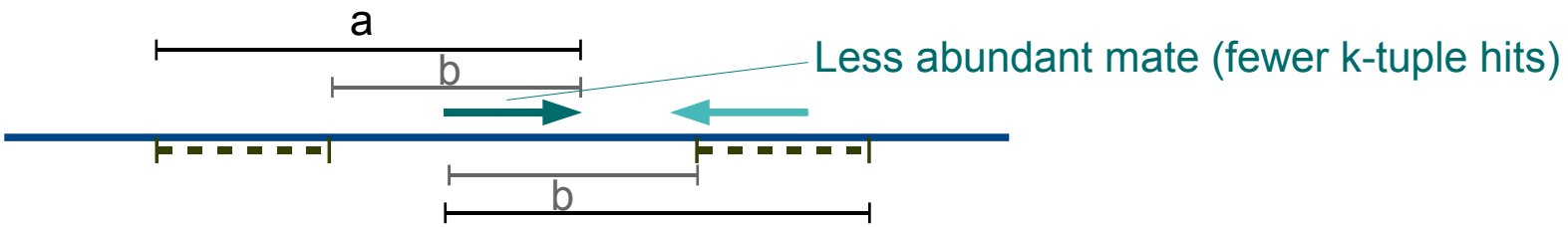


iv) Report both mates if the more abundant mate maps uniquely in the distance band [a,b] around the matching positions of the less abundant mate.

wellcome trust sanger institute

Module for paired-end reads

Command line options: *-pair <a,b> [-outfile <filnam>]*



iv) Report both mates if the more abundant mate maps uniquely in the distance band [a,b] around the matching positions of the less abundant mate.

Otherwise:

va) if both mates map uniquely (mapping score > 30) outside the specified distance range → report pair to a separate file (specified with *-outfile <filnam>*)

vb) if only one mate maps uniquely, report this single mate

wellcome trust sanger institute

Command line examples

> ssaha2 -rtype solexa -pair 180,220 -kmer 12 -skip 2 -output cigar -align 1 genome.fa paired_end_reads.fq

Or build hash table on big memory machine

> ssaha2Build -kmer 13 -skip 2 -save htab genome.fa

and run ssaha2 with reduced memory usage (farm):

> ssaha2 -rtype solexa -pair 50,600 -outfile pairs.out -disk 1 -output cigar -save htab paired_end_reads.fq

wellcome trust sanger institute

The SSAHA2 package

Binaries available from:
http://www.sanger.ac.uk/Software/analysis/SSAHA2

Acknowledgements

Zemin Ning

Tony Cox

Yong Gu

Adam Spargo

Ben Blackburne

wellcome trust sanger institute

Current SSAHA2 performance with short reads

wellcome trust sanger institute