

Assembling Sequences Using Trace Signals and Additional Sequence Information

Bastien Chevreux, Thomas Pfisterer, Thomas Wetter, Sandor Suhai

Deutsches Krebsforschungszentrum Heidelberg akiz

## **Problem definition**

























**T**Z

### Signal problems





## **DNA problems**

Chemical properties

Coiling of DNA
Problems with dye chemistry

Repetitive elements

Standard short term repeat (ALU, REPT etc.)
Long term repeats of sometimes several kb



### **Conventional assembly**





### Integrated Assembler-Editor



## **Assembler: Input**

- Collection of reads
  - unknown relationship
  - unknown direction

#### Each read

- unknown error distribution
- sequencing vector tagged
- trace signal information
- opt. base quality values
- opt. quality clipping, marking HCRs (High Confidence Regions)
- opt. standard repeats tagged
- opt. template information



## **Assembly: Framework**

- Establishing relationships of each read against each other results in full oversight over the whole assembly
- Problem: k reads -> time complexity O(k<sup>2</sup>)
- Fast read comparison routines needed
- Smith-Waterman has O(mn), very slow



## **DNA-SAND** algorithm

- Shift-AND algorithm: fault tolerant, O(cmn)
- modified Shift-AND for read comparison, DNA-SAND: fault tolerant, O(cn) with 0<c<12</li>
- high sensitivity and specificity

   less than 0.75% missed overlaps
   around 45-50% false positive hits



## **Assembly: Framework**

#### Fault tolerant

- Sandsieve principle: obvious mismatches discarded, potential matches remembered
- Check each read in forward and reverse complement direction

		- 14 - 14 - 14 - 14 - 14 - 14 - 14 - 14	1	1			10.1		1		-				1.1				4
			ي المان		- 19 July 1		1.1.2			5 - SP							1.1	11	
		1711				101	1.11		3.50	11	1.1	. :	111				1.1		
	e de Maria	.42.50	. z: .;	1 I I I			1	an 6.		÷ 3.	6 A		÷ ::				11.1		
-	1.27	n este en	ં છે. ં હ	0.4	ليحقب ا	ę. il i	36 . SG	್ಯ ನಿ	1.1.22		i y i	d sy f	14.17	èn t	Р. v	1.11	- C.	10.00	1.1
- A.	1.11.4	56E (*	- <u>1</u>		1000	1.2.6	hi ch	200	-915	i a se			1913	60.	- S. 1.	- 1		- <u>`</u>	
- A.	e si ingi	97 <b>2</b> ()	é É le pie		الشيونية.	na na ra					uru'- i	8. G.F	e la companya da serie da s Serie da serie	4. A ( )		- 1 - P	al, al	5.00	л÷
1.1		61167 13	이 바이 바이 바이 바이 아이 아이에 가지 않는다. 이 아이에 아이에 아이에 아이에 아이에 아이에 아이에 아이에 아이에 아이		341-5541	¢, n ⊫s	이다. 나라네 - ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	9 <b>1 6</b> 15 3		C1 - (81-	100110	Ni: itik Ali taniz	고문	811 (J.	1022	6 B	10.00	191	11
1114		े हे हैं।	- <b>B</b> - (\$		it	1 A -	96,92 P	d 14	8.00	heljio.	eren i	6,57	g R	6 N S	1.1		110		1
	1.0.1112.116	en el 18 de la	e en el sub	19.4.18-42		ST 18 (12	065614	ia sir y	0 W K	UH ( 501 )	i na s	hr thir	ai in	$800 \le l$	$E \in \mathbb{N}$	- 1 <b>2</b> 1	149-09-	. # Č	< 3
1.1	S. 84.	1111	1990 - 1990 1980 - 1990	a de la com			1.12.1	400	1.1				말하네.	12.1	1 A .	· · ·	1.1.1.1	- <u>1</u> 1-	
1.12	: a ::h	ម <b>ខេត្</b> ម ខ	ି ଛା ୁ ୁ କ	n an àirte	. esti i Heau	necind ru	o dia Au	usta is a	piare e	n bai	i je pu	ar are	- igae	9-0 m	2.35		s estern	19.5	1
			22128	1.1	1211				- 11 - L			1 V				1.1		· · .	
		나버린을	2 DI 194	10.132	444	- <u></u>	، <b>م</b> ر فدادیا	di seren di Seren di seren di sere	كالإستث	ka jeri i	- 34 - 5	il seri	a de la com	a i bi	1.10	1.11	11.2		
	12.1									· •							1.1		
1.14	ाः जुष्	244 S	. k	9	1 di 13	er gi, ir	ier a		31	6.52	i i gi t	é	10.0		1	. °-	1.1		
14 - 1 <sup>0</sup>	· · · · · · · · · · · · · · · · · · ·	20 F. # 4 -	<b>1</b> 51	영국학교		20.1	900 L 24 M	02.2 011	196,521		nun e	탄 산송	t ti tir	981 I	$\mathcal{T} \geq \mathcal{T}$	y et da	t d'ange	t ditt.	1
		1.2		141.				44	1412	15791		t. 1. l.	105	1. J. J.	1.51		1.0		
en j	요구분명	18	의 범인구별	관관금	一种新闻	<b>新日</b> 日	en fiere	14056	경원	医测定	지난	60.Hel(	s (di lit	81 : :	-1344	1.44	13 e I	200	ê9)
÷ .'	t, naisi	七 타팔 감종	é) an t-Sa	Fig. Are	. HØ15.86	動に創いの	ani: Can	्य आहे व	នាំងដែរ	E (観)	tata a	ត្) ដោ	- <del>14</del> 6	60 M	$\{ e_i, i \in$	1.4	$1 \le n \le 1$	1 - 181) -	Ġ.
-1.5	e di figel e g	1.0	81 b	्रेक्ट	1941 - 55	the ta_t	112-11	، ار بارد،	e 174 -	اجراد م	1.87.5	$a_1 + a_2 + c_3$	-iuia	ies 41	100	an di	11.1		14
111.0		日日書名	빈 분 한 2 분	日日二				-8482	영감	101-	n per e	5.25	313	n p	11日	1, 61,	일문	- 20.	11
	11.11.1	921	1 M L D	2 C.	- 373	6 H L	813-7	1.1	2191	F. 1917			9 t. d.		1.021	4 E	11.14	1.1	
1.1		1111			1.27	5 101	117.1		1.41.54				1.2	0.43	1. 1		1.1	1.1	
		1121							2.24.)	- 1		4. T	-99	24.54	1.10		· · · ·	- file	11
11.1	11111	1. 21	- S - S	1.70	· 1000	8.8	10 a 200	1			123		영문	26.63	1.11	1	· · ·	- C.	1.1
		1.4.4	61	12 E F	1550.52	fe ic di	90 i < 4	a, b	1.18	6 (R)	( bi s	e det	din 1	e da	her.	ð., e	194	14	£
-11-1			9-5 y		1.7.8.22	2.2.3				L						· ·		1.000	·
		0120	18 V.		1.50	a. e. j	0.123	2.90	- 2.4	2.40			1.1	1. 1			1.11		
	n koʻzmi		E H	e ji s	2012-01	n 18, E	-1+1 > 8+1	100		8 E C	5 de 5.1	é tri f	$2\pi \dot{\varphi}$	211 J	1972-1	··	- 영국	1.2	
· .	11 July 10		e. 80	111	- 2월 문화	1		2.22	Contraction (	an prin	신문		212.0	1.00	11.		3.1	. 31	
			- B	1.27	1.1.1	6 P 1				1.56	140 J			et. 19	1.11	12		- N	Ε.
	to Asaba	いたますの	6 第1日前常	10.311	- 김태영영	8 8 C	nitasin	्रहेल्ल	School	ensi.	88	क्षा राजे	$dd_{12}$	en e	ger fri		11 2.4		- 3
		and the second second		25.1	بالم الم		949 J.	ong:	1.20			181			1.1.	1.1	1.1	<sup>1</sup> .	
11. J.	11. J. 1	1.24.01	- 21		1.111-13	1,42,52	(R. 1	46.24				42.45	i ka d	24 (P	4.5	73	- St. 1	1.1.1	17
	a sa ta As		ے پر اندر ت	الد من					5425		. 21.			с. 23.		1.1		· .	
	1.01.1	1122		1.1.1	1919.3	220	· · · · · · · · · · · · · · · · · · ·	4123	1143		111	211	1.1		· · · ·	- E.	161	1.1.1	
	11 de 2	-1 <del>-</del> 1 -	이 같은 것이 같은 것이 같이 같이 같이 같이 했다.		a develo	s err	1	공연		- ÷	e Ma	4,159	- 1 e	e de la	1.1	1.11		11.1	
	1.1.1.1.1	1. J	11.1				an a	100	1.1		224	S. A.	:: ' î î		2		1.13	1.1	. 1
			e elg je	en é -	· · · · · · ·		antan A				11 - 11 - 11 - 11 11 - 11 - 11 - 11 - 1	1. C.	ing ny	· .	. *	· •			
	1 - 19 - 19 - 19 - 19 - 19 - 19 - 19 - 1			13 T F	- 11 C - 23			1.5		E 81			. 11 1	12.14			- 1. C.	1.1	. î
		4382		C (A)	in the second	S. 845.	11 A.	1211	1,143	÷ 1.	1217		1.1		· ·		1.1		
		199	- C		100	e de la com	1.55		S			19 G				· ,			
	e and d	. i.e	i ka sa sa	'.	A stations		£				. i i					2			
		11. L.	1.1	- 11 - E	100	1 M L	99 J. 1	7.1									S	11.1	
10.0		n na gi	. <u>1</u> . 1	n de e	- N (19)	다 나라	en e	p, i	1947	1 . M	28.17	5 J. 25			(1, 1)		day -	- A.,	
	1. 1. 1. 1. 1.	1995)	144.0	11 M.C	- 1, Q.C. (c.	S Die	42 分。	21.	1.0	E, 2.,	- 6.5	9 J.	100	ст <u>.</u> н	100	$M_{\rm eff} = M_{\rm eff}$	- N	<sup>1</sup> .	
	1.01	ça de	6 A. (1)	1.12	1.1	÷.,	1.1		- i - i -		÷.,		2.11	197		1.00	•	14. 1	
11		$-e^{2\beta}$	39 S		÷ - ÷	e gener	9.551 2	$(q_{1}, p_{2})$	- 22.1	( . R.	Aq e i	<u>1</u> . 19	1.1.2	1.11			- 11	2	
		- 17 A)	100	1. J. C. S.		2,214	din 18	14 a.,		. e	1.0	) e	- <sup>26</sup> - 4	÷.,	1.25		11	1	



## **Overlap confirmation**

- Evaluates potential overlaps
- Standard (banded) Smith-Waterman algorithm: max(O(bm), O(bn))
- Rough calculation of SW match quality, eliminating false positive DNA-SAND matches
- Calculate an "alignment weight" for accepted overlaps



### **Overlap confirmation**



- Accepted match
  - Overlap: 196
     bases
  - Score: 180
  - Score ratio: 92%
- Weight: 151817



- Rejected match
  - Out of band!
  - Overlap: 204 bases
  - Score: 133
  - Score ratio: 65%



## Building a weighted graph

Example: 6 reads



All possible overlaps for 2 reads



## Building a weighted graph

1

Pruned by DNA-SAND





# Building a weighted graph





# **Building contigs**

- Multiple alignment is too slow
- Building a consensus by iteratively aligning reads against existing consensus
- Important:
  - Order of read alignments
  - Finding good alignment candidates
  - Possibility to reject candidates

## Interaction: Pathfinder & Contig

#### Pathfinder:

- search good starting point for contig building
- find good alignment candidates to add to existing contig
- always inspect alternative paths in overlap graph

#### Contig:

- accept reads that match to existing consensus
- reject reads that do not match
- find inconsistencies that 'build up slowly' and mark these



## Pathfinder: Strategy

Finding starting points:
Search for node with a high number of reasonably weighted edges
Exclude edges below threshold
Finding next alignment candidate:
Find reads with best nodes in contig
Recursively analyse best edges in graph



# Contig: Strategy

- Align given read of given edge to existing contig at approximated position
- Accept read that match
- Reject reads that introduce
  - significantly higher error rates in contig than predicted by weighted edge
  - many non-editable errors in repetitive regions
  - inconsistencies with given template insert sizes

## **Contig: Raw**



GKZ

# **Contig: Edited**

0372

![](_page_26_Figure_1.jpeg)

## **Contig: Raw**

![](_page_27_Figure_1.jpeg)

Contig:U13a09h09.t1(#416) Length:22938

0 KHZ

## **Contig: Edited**

0 K 7

![](_page_28_Figure_1.jpeg)

![](_page_29_Picture_0.jpeg)

	Contig Editor: -435	U13a04	d08.t1									一河
< C:	1 > < Q: -	1 >	🗆 Insert	Edit Modes	Cutoffs	Undo Next Search	Commands S	ettings			Quit	Help
<<	< >	>>									Þ	Ζ
		$\geq$	2	310	2320	2330	2340	2350	2360	2370	238	0
406	U13a06d03.t1		AAAGTAC	:TAT <mark>-</mark> AGAA	TAATTTA	ITA <mark>G×</mark> CAATGAG1	ттаааастс	AACTTTTTAT	[GCCTTC*TCf	ICCAGCTGCAA	AGTGT	T
407	U13e07f02.t2		AAAGTAC	TATAAGAA	TAATTTA	ITA <mark>G×</mark> CAATGAGT	ттаааастс	GACTTTTTATI	igccttc*tcf	ACCAGCTGCAA	AGTGT	T
411	U13a05b10.t1		AAAGTAC	TATAAGAA	TAATTTA	ITA <mark>G</mark> GCAA <mark>G</mark> GAGT	ттаааастс	ACTITITAT	IGCCTTC*TCF	ACC <mark>-</mark> GCTGCAA	AGTGT	T
405	U13a08f11.t1		AAAGTAC	:TA <mark>-</mark> AAGAA	TAATTTA	ITA¥GCAATGAG1	T <mark>-</mark> AAAACTC	AACTTTTTAT	GCCTTC*TCf	ACCA <mark>-</mark> CTGCAA	AGTGT	T
404	U13a09h09.t1		AAAGTAC	TATAAGAA	TAATTTA	ITA*GCAATGAG1	TTAAAACT	AACTTTTTATI	IGCCTTC <mark>A</mark> TCA	nc <mark>-</mark> agctgcaa	AGTGT	T
331	U13a03g02.t1		AAAGTAC	TATAAGAA	TAATTTA	ita*g <mark>c</mark> aatgag1	ttaaaa <mark>c</mark> tc	AACTTTTTAT	IGCCTTC <del>*</del> TCA	NCCAGCTGCAA	AGTGT	T
-330	U13e07d02.t2			AGAA	TAATTTA	IT <mark>=</mark> *G <mark>C</mark> AATG <mark>a</mark> GT	'TTAAAA <mark>C</mark> TC	AActititat	:GC <mark>c</mark> TTC*TCf	ACCAGCTGC <mark>×</mark> A	AGTGT	T
-335	U13a04c06.t1			AA	TAATTTA	ITA*G <mark>C</mark> AATGAG1	'TTAAAA <mark>C</mark> TC	AACTTTTTAT	GCCTTC*TCf	ACCAGCTGCAA	AGTGT	T
	CONSENSUS		AAAGTAC	TATAAGAA	TAATTTA	ITA*G <mark>C</mark> AATGAGT	ttaaaa <mark>c</mark> tc	AACTTTTTAT	GCCTTC*TCA	ICCAGCTGCAA	AGTGT	T
Tag type	Tag type:PRMB Direction:+ Comment:"Possible Repeat Marker Base found by MIRA."											

	Contig Editor: -435	U13a04	d08.t1								-124	
< C:	1 > < Q: -1	>	🗆 Insert 🛛 E	dit Modes 🔲 Cut	offs Undo Next S	Search Command	s Settings			Quit H	lelp	
<<	< >	>>		1						$\geq$	Ĩ	
		$\geq$	2300	2310	2320	2330	2340	2350	2360	2370		
406	U13a06d03.t1		AAAGTACT	at <mark>a</mark> agaataat	TTATAGCAATG	AGTTTAAAAC	TCAACTTTTT	ATTGCCTTCT	CACCAGCTGC	AAAGTGTTTT	1	
407	U13e07f02.t2		AAAGTACT	at <mark>a</mark> agaataat	TTATAGCAATG	AGTTTAAAAC	tc <mark>g</mark> acttttt	ATTGCCTTCT	CACCAGCTGC	AAAGTGTTTT	1	
411	U13a05b10.t1		AAAGTACT	ataagaataa1	TTAT <mark>A</mark> GCAA <mark>t</mark> G	AGTTTAAAAC	TCAACTTTTT	ATTGCCTTCT	CACC <mark>a</mark> GCTGC	AAAGTGTTT		
405	U13a08f11.t1		AAAGTACT	"A <mark>t</mark> AAGAATAA1	TTATAGCAATG	iagtt <mark>t</mark> aaaac`	TCAACTTTTT	ATTGCCTTCT	CACCAgCTGC	AAAGTGTTTT	1	
404	U13a09h09.t1		AAAGTACT	ataagaataa1	TTATAGCAATG	agtttaaaac	T <mark>-</mark> AACTTTTT	ATTGCCTT <mark>C</mark> T	CAC <mark>c</mark> AGCTGC	AAAGTGTTTT	1	
331	U13a03g02.t1		AAAGTACT	ATAAGAATAA1	TTATAG <mark>C</mark> AATG	agtttaaaa <mark>c</mark>	TCAACTTTTT	ATTGCCTTCT	CACCAGCTGC	AAAGTGTTTT	1	
-330	U13e07d02.t2			AGAATAAT	TTAT <mark>a</mark> G <mark>C</mark> AATG	agtttaaaa <mark>c</mark>	tcaa <mark>c</mark> ttttt	AT <mark>t</mark> GC <mark>c</mark> TTCT	CACCAGCTGC	* <mark>AAGTGTTTT</mark>	1	
-335	U13a04c06.t1			AATAAT	TTATAG <mark>C</mark> AATG	agtttaaaa <mark>c</mark>	TCAACTTTTT	AT <mark>t</mark> GCCTTCT	CACCAGCTGC	AAAGTGTTTT	1	
	CONSENSUS		AAAGTACT	ATAAGAATAAT	TTATAG <mark>C</mark> AATG	agtttaaaac	TCAACTTTTT	ATTGCCTTC <mark>T</mark>	CACCAGCTGC	AAAGTGTTTT	T	
Tag type	Tag type:PRMB Direction:- Comment:"Possible Repeat Marker Base found by MIRA."											

![](_page_30_Picture_0.jpeg)

	Contig Editor: -98	8 U13a04q	j12.t1									-14
< C:	1 > < Q: -	-1 >	💷 Insert 🛛 E	dit Modes 🔲 C	utoffs Undo Ne	ext Search Com	mands Settings				Quit	Help
<<	< >	>>		]							Þ	Σ
		$\geq$	1740	1750	1760	1770	1780	1790	1800	1810		
-80	U13a06b01.t1	1	AGACACCA	AAATAGTGG	ATGCCCAAGT	CCTGATAGA	AAAAATTGCAT	AGTATCTGC	ATATGACCTA	TGCAATCC	тсст	G
81	U13a02e11.t1	1	AGACACCA	AAATAGTGG	ATGCCC <mark>a</mark> AGT(	CCTGATAGA	AAAAATTGC <mark>a</mark>					
104	U13a06h05.t1	1	AgACACCA	ì								
-79	U13a01e08.t1	1	AGACACCA	AAATAGTGG	ATGCCCAAGT	CCTGATAGA	AAAAATTGCAT	AGTATCTGC	ATATGACCTA	TGCAATCC	тсст	G
111	U13a02c01.t1	1	AGACACCA	AAATAGTGG	AT							
78	U13a07b08.t1	1	AGACACCA	AAATAGTGG	ATGCCCAAGT	CCTGATAGA	AAAAATTGCAT	AGTATCTGC	ATATGACCTA	TGCAATCC	тсст	G
-90	U13a06d12.t1	1	AGACACCA	AAATAGTGG	ATGCCCAAGT	CCTGATAGA	AAAAATTGCAT	AGTATCTGC	ATATGACCTA	TGCAATCC	тсст	G
-101	U13a08b12.t1	1								C	тсст	G
	CONSENSUS		AGACACCA	AAATAGTGG	ATGCCCAAGT	CCTGATAGA	AAAAATTGCAT	AGTATCTGC	ATATGACCTA	TGCAATCC	тсст	G 🔽
Tag type	Tag type:ED_C Direction:+ Comment:"Alter Base at 1747 1 # N g U13a06h05.t1 287"											

X X X	Contig Editor: -98 U13a04	j12.t1							-14	
< C:	1 > < Q: -1 >	💷 Insert Edit Modes	Cutoffs Undo	Next Search	Commands Settings			Quit	Help	
<<	< > >>								Ζ	
$\overline{\ }$		1740 1750	1760	1770	1780	1790	1800	1810		
-80	U13a06b01.t1	AGACACCAAAATAG	FGGATGCCCAA	GTCCCTGAT	AGAAAAAATTGCAT	AGTATCTGCAT	ATGACCTATE	CAATCCTCCT	G	
81	U13a02e11.t1	AGACACCAAAATAG	iggatgccc <mark>a</mark> a(	GTCCCTGAT	agaaaaaattgc <mark>a</mark> ti	AGTATCTGCAT	ATGACCTATE	CAATCCTCCT(	G	
104	U13a06h05.t1	Agacaccaaaatag	rggatgccc <mark>a</mark> a(	GTCCC <mark>-</mark> GAT	AGAAAAAATTGCAT	AGTATCTGCAT	ATGACCTATE	CAATCCTCCT(	G	
-79	U13a01e08.t1	AGACACCAAAATAG	FGGATGCCCAA	GTCCCTGAT	AGAAAAAATTGCAT	AGTATCTGCAT	TATGACCTATE	CAATCCTCCT	G	
111	U13a02c01.t1	AGACACCAAAATAG	<b>FGGAT</b> GCCCAA	GTCCC <mark>-</mark> GAT	AGAAAAAATTGCAT	AGT <mark>-</mark> TCTGCAT	<b>ATGACCTATE</b>	CAATCCTCCT(	G	
78	U13a07b08.t1	AGACACCAAAATAG	FGGATGCCCAA	GTCCCTGAT	AGAAAAAATTGCAT	AGTATCTGCAT	TATGACCTATE	CAATCCTCCT	G	
-90	U13a06d12.t1	AGACACCAAAATAG	FGGATGCCCAA	GTCCCTGAT	AGAAAAAATTGCAT	AGTATCTGCAT	TATGACCTATE	CAATCCTCCT	G	
-101	U13a08b12.t1	AGACACCAAAATAG	FGG <mark>-</mark> TGCCCAA(	GTCC <mark>T</mark> TGAT	AG <mark>-</mark> AAAA <mark>-</mark> TTGCAT	AGTAT <mark>G-</mark> GCAT	TATGACCTA <mark>-</mark> G	ICAATCCTCCT	G	
-107	U13a01a07.t1	CAGACA <mark>CC</mark> AAA <mark>ATA</mark>	G <mark>-</mark> G <mark>G</mark> TGCCCAA(	GTCC <u>C</u> TGAT	AGAAAAAAT <mark>G</mark> GCATI	AGTAT <mark>-</mark> TGCAT	a <mark>g</mark> gacctate	CAATCCTCCT(	G	
-75	U13a04f02.t1							C <mark>A</mark>	C	
-131	U13a010c08.t1	Ga <mark>aa</mark> a-GGTGGG <mark>a</mark> T·	-GG <mark>T</mark> T <mark>AAAA</mark> A-(	AG———GC <mark>A</mark> A	-TGGGGG- <mark>T</mark> AAGGG	G <mark>G</mark> G <mark>A</mark> AA-AA	- <mark>at</mark> agaa-caa	iagg-aaaagai	A	
-125	U13a08b01.t1	-T-TT <mark>CC</mark> CGTA-AA	GAT <mark>G</mark> A-G-T	rcgagat <u></u> ga	-C <mark>a-CCTC</mark> TTCACA	- <u>C</u> a <mark>a</mark> -gaaggo	ì−AT <mark>A</mark> −A−G−C	C <mark>AAGG-AA</mark>	G	
	CONSENSUS	AGACACCAAAATAG	GGATGCCCAA	GTCCCTGAT	AGAAAAAATTGCAT	AGTATCTGCAT	ATGACCTATO	CAATCCTCCT	G	
Tag type:ED_C Direction:+ Comment:"Alter Base at 1747 1 # N g U13a06h05.t1 287"										

![](_page_31_Picture_0.jpeg)

## **Extending HCRs**

- ´beef up´existing contigs; trivial, very fast
- extend existing contigs; simple, quick
- find new contigs to build; bold, slow

![](_page_32_Picture_0.jpeg)

Data preprocessing

#### Fast read compariso<mark>n</mark>

![](_page_32_Figure_3.jpeg)

#### **Read extension**

Pathfinder

#### Repeat marker Automatic editing

**Contig assembly** 

**Finished project** 

#### azz

### Status

- beta-testing almost completed
- assembler & editor in use to assemble projects up to 10.000 reads
- first evaluation: human finished 35kb project (Golden Standard) without fine-tuning assembled contigs have 99,9x% identity
- whole genome shotgun with 23.000 reads in preparation
- other applications like EST clustering?

![](_page_34_Picture_0.jpeg)

## **Canonical Homepage**

#### http://www.dkfz-heidelberg.de/mbp-ased/

## Acknowledgements

Prof. Rosenthal, Matthias Platzer, Uwe Menzel and the IMB Jena genome sequencing centre Bernd Drescher and Lion Biosciences AG, Heidelberg