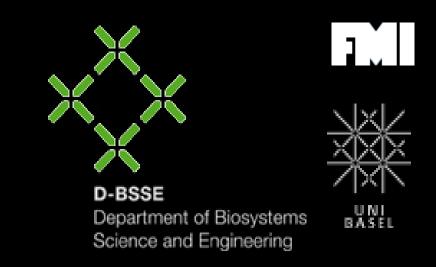
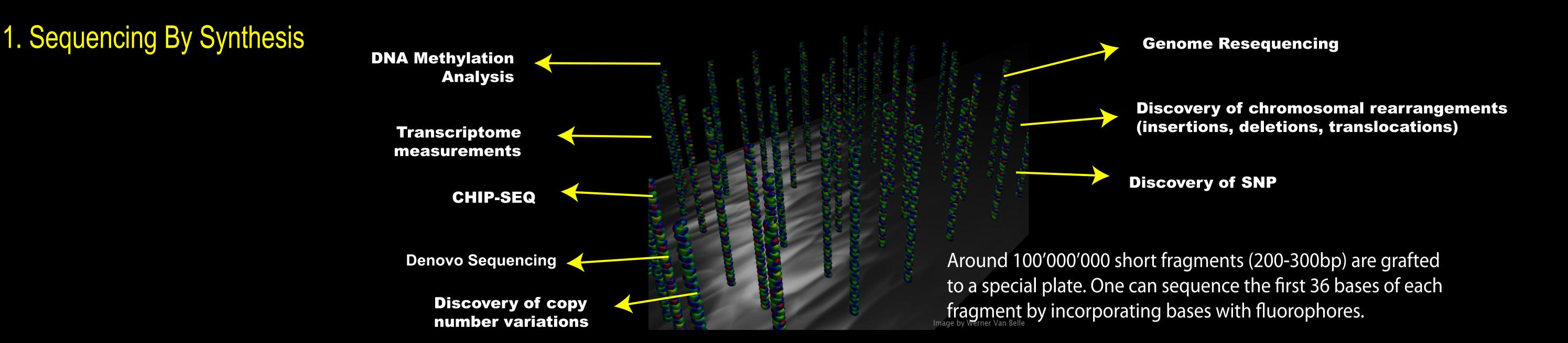
Deep Sequencing



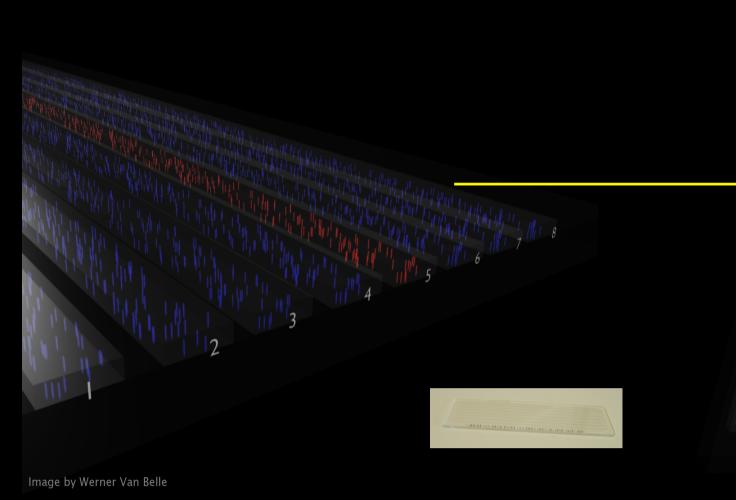
Werner Van Belle, Ina Nissen, Christian Beisel

Laboratory for Quantitative Genomics

Department Biosystems Science and Engineering - ETH Zurich

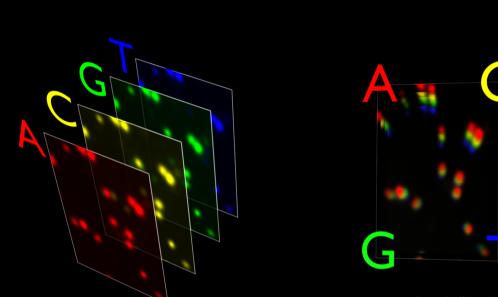


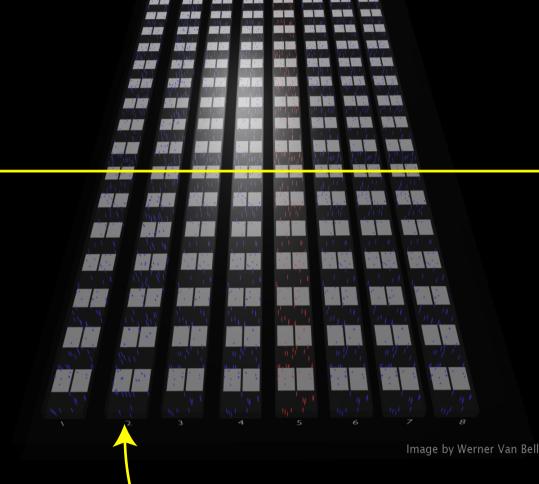
2. Image Acquisition

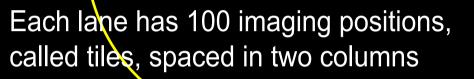


A flowcell consists of 8 lanes. Each lane can contain one biological sample. Lane 5 is always reserved for a control sample (PhiX).

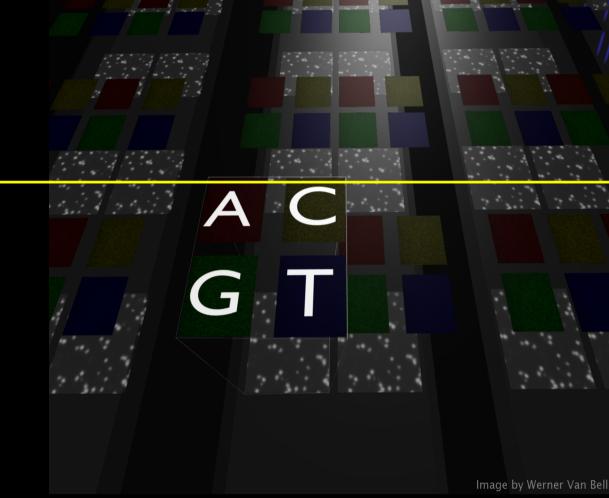






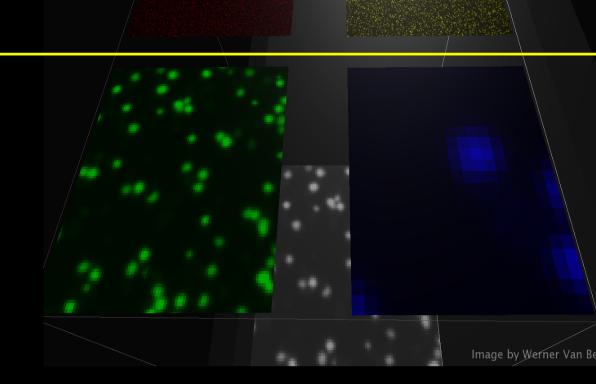


4. Basecalling



Each tile is imaged 4 times. Once for each fluorophore.

36 x



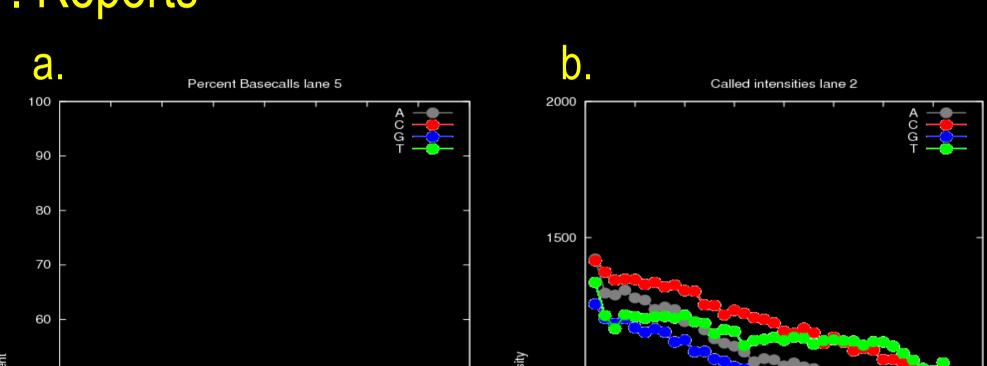
4 tile images. Topleft: normal size. Topright: 2x Bottomleft: 50x. Bottomright 100x zoom. Each spot is a short fragment (cluster)

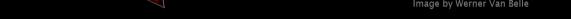
The imaging cycle is repeated 36 times. Between two imaging cycles we incorporate a new base.

Image by Werner Va

0

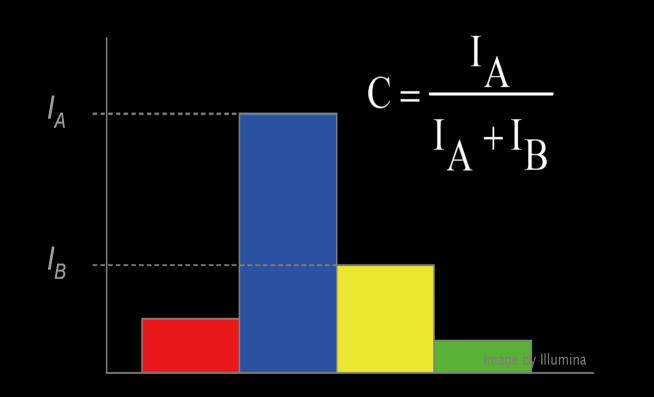






Because the lightpath for each of the filters is slightly different we must first align all images. Afterwards clusters are identified and listed. This is done with Firecrest.

5. Data Filtering



Further data filtering is based on the ratio between the highest intensity and the second highest intensity. If the ratio C is smaller than 0.6 then the base can be considered unclean.

Each cluster has 36 intensities for A, C, G and T. These intensities require further normalization because a) certain molecules will run ahead or lag behind (phasing and pre-phasing) and b) there is crosstalk between the A, C, G and T channels. Afterwards, the basecalling can be performed. This is done with Bustard.

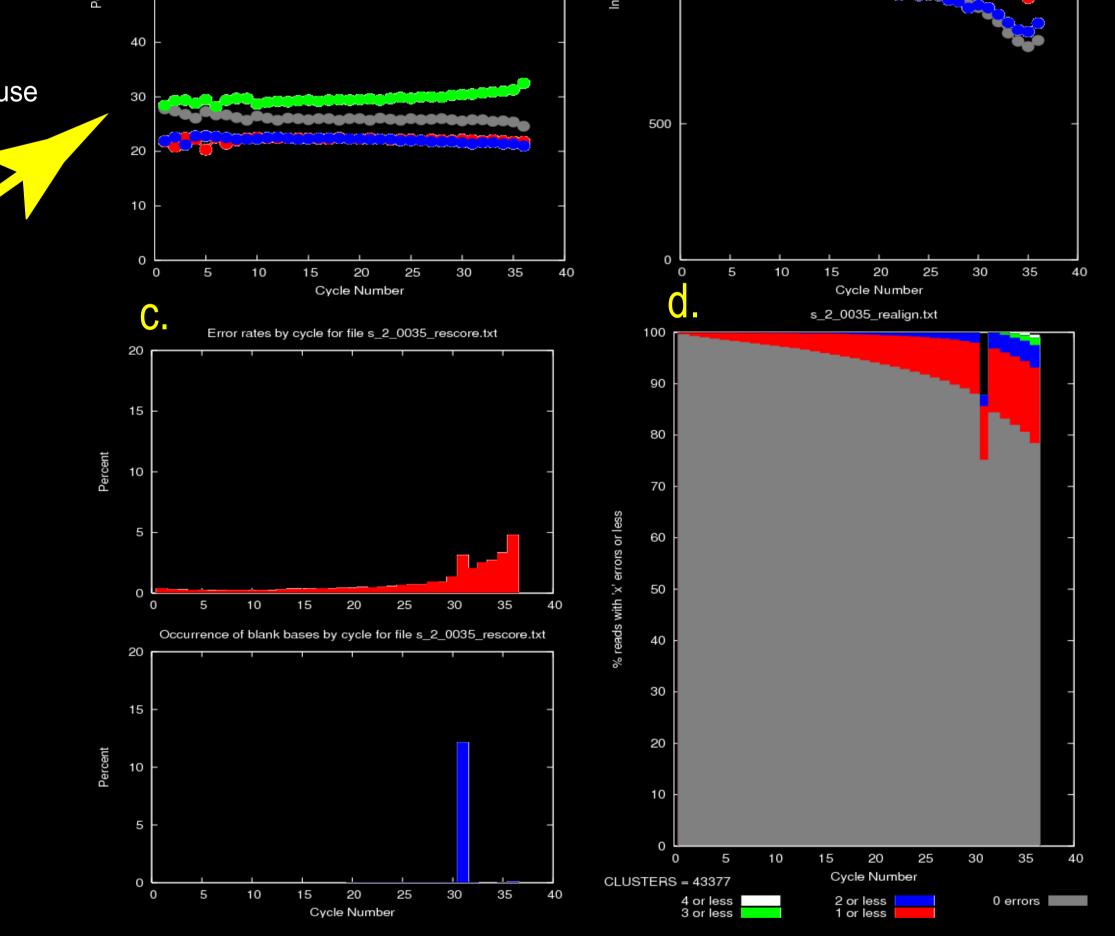
A

6. Short Fragment Alignment

GAATAAGGTGAAAAGCTTTGATTTGCCT ACATCTCGAAGACGGTAAGACAGTCTGC CCAGTCTGTAGT<mark>T</mark>AATAAGGTGAA<mark>T</mark>AGATTTGGTT ACGCCGCGCGGACCTCTGGAAGACGGTAAGACTAT GTTTGCCTCAAACATCCCAGACGCCGCC CTCCCAG<mark>G</mark>CTGTAGTGAATAAGGTGAAAA<mark>A</mark>ATTTG GTAT**T**AGAGCTTTGGGCATTCGCA IGGTTTACCTCAAACATCCCAGACGACGCGCGGAC

GCGTACCTGTGGTATAAGAG Image by Werner Van Belle

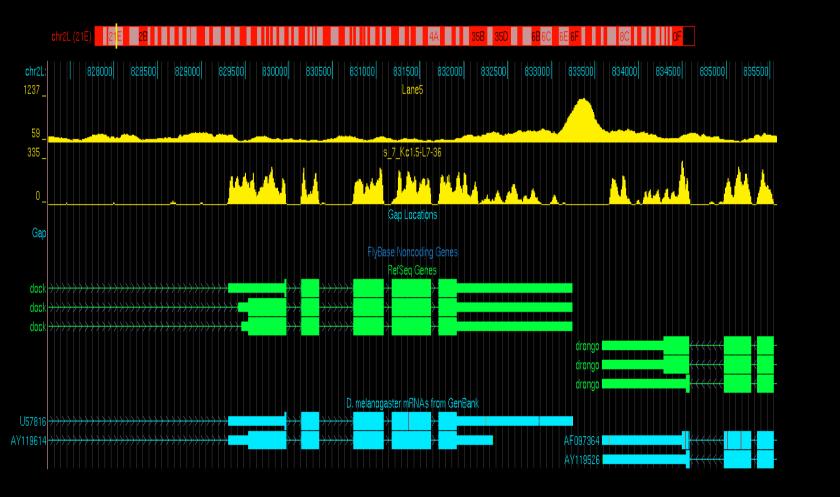
The short fragments can be aligned to a reference genome (yellow). The first program (Eland) allows for at most two mismatches per fragment. The second program (PhageAlign) will find the best match for each fragment.



Gerald generates various error reports a) % basecalls per base per cycle b) average intensities for each base per cycle c) statistics on base mismatches per cycle d) % of read errors per base per cycle



8. Genome Browsing



After converting the alignment files to WIG files one can visualize them in the UCSC genome browser. Track 1) output from a CHIP-SEQ experiment. Track 2) output from an RNA experiment. Notice the sharp exon boundaries aligning perfectly to the genome tracks.

9. Expression Reports



\$	A	В	C	D	E	F	G	H	I	J	K	L	M	N	0	Р	Q	E
1	Exon		Intron		Expression	0.961517222												
2	Mass	Length	Mass	Length	Total	Exon	CG	Descriptio	n									
3	3309228	814	304795	183	3624.897693	4065.390663	CG9282	60S riboson	nal protein L24	. [Source:L	Iniprot/SWI	SSPROT; Ad	cc:Q9VJY6]					
4	2776390	798	748760	965	1999.517867	3479.185464	CG7808	40S riboson	nal protein S8.	[Source:Ur	niprot/SWIS	SPROT; Acc	c:Q8MLY8]					
5	3113478	1011	280139	448	2325.988348	3079.602374	CG11522	Ribosomal p	protein L6 CG1	- 1522-PB, is	oform B [So	urce:RefS	eq_peptide;Ac	c:NP_6518	76]			
6	2726923	908	162726	265	2463.468883	3003.219163	CG6779		nal protein S3.									
7	4340876	1458	499825	672	2272.629577	2977.281207	CG5502		nal protein L4 (-			, ,					i l
8	7535847	2676				2816.086323			70 kDa protein	. , L			· -		orot/SWISSPR	OT:Acc:P111	471	i li
9	5002990	1832				2730.889738			na-1 chain. [So				, L					
10	1953664	723				2702.163209			nal protein S29				-					
11	1995472	739				2700.232747			nal protein L13				<u> </u>					
12	1540676	576				2674.784722			protein S30 CG					cc:NP 732	5661			
13	2251277	844				2667.38981			protein L27 CG									
14	1930520	736				2622.98913			nal protein S25									
15	3039445	1317				2307.854973			nal protein S7.	-			-					
16	892566	403				2214.80397			prant-binding p				· ·		V8Y91			
17	2181894	1031				2116.28904			ally-controlled t							521		
18	1268885	675				1879.82963			protein L35A C						,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,			
19	2227098	1204		1		1849.749169			[Source:RefSe					555				
20	1542457	856				1801.935748			nal protein L32			-	·Uninrot/SWIS	SPROTIACC	·P043591			
21	974040	562				1733.16726			nal protein S20			, L						
22	2965406	1752				1692.583333			factor 1-gamm	-				prot/SWISS	SPROTIACCIOS			
23	1692571	1016				1665.916339			nal protein S3a		, ,	-	, L		Si Kor, Acc. QJ			
24	2088834	1261				1656.490087			nd still. [Source	· ·	, L							
25	1772259	1086				1631.914365			[Source:RefSe				2]					
26	1149387	715				1607.534266			protein S27 CG			-	Acc:NP 651	3501				
27	1365852	881				1550.342792			9354-PA, isofor									
28	3280145	2441				1343.770995			nal protein L3.				NP_049887]					
29	2528588	1914				1321.101358			A [Source:RefS	-								
30	1839192	1914				1257.137389			precursor (CRF									
31	1597329	1463				1080.736806				, ,		10 🗕						
32	1492361	1478				1040.698047			ein epsilon (Su									
33	1172022								A [Source:RefS									100
33 34		1167				1004.303342			RC CG6378-PA			1 +						
35	672760	681				987.9001468			diphosphate ki							ندر ×	1	
35 36	863989	897				963.1984392			A [Source:RefS							A STATE OF		
36 37	838186 3721111	906				925.1501104			A variant. [Sou			0.1 +				and the second second		
		4267				872.0672604			A [Source:RefS							1		
38	3800526	4407				862.3839346			[Source:RefSe						A A MERICAN	all a		
39	986215	1210				815.053719			S-transferase			0.01 +		× × ×		× ×		
40	5687806	7059				805.7523729		CG14967-P	A [Source:RefS	eq_peptide	;ACC:NP_64			A A A A A A A A A A A A A A A A A A A	1			
41	4995902	6262				797.81252		C			tida : A NR							
42	1827849	2353				776.8164046			2-PB [Source:F			0.001 +						
43	505664	656				770.8292683			A [Source:RefS				××	Pitto	R 10			
44	1208685	1600				755.428125		CG17746-P	B, isoform B [S	ource:RefS	eq_peptide		× x*× 🖓					
45	520710	694				750.3025937						0.0001 +	× ****	*				
46	536165	720				744.6736111												
47	648988	873	819			743.3997709	CG18179	CG18179-P	A [Source:RefS	eq_peptide	;Acc:NP_64							
48		2004 Sheet1 +	57154	114	717 8985507	720 6876701	CG32351	CG33351_D	A [Source: Pef9	ea nentida	Acc: NID 72	0.00001 +						
		Peady				Sum_0		O SCRI O CAR				0.000	04 0.0004	0.004	0.01	0.4	4	10

By counting the number of bases matched at a certain genome position, we can accurately report the expression level of each gene. This includes exon and intron expressions. The correlation between two technical replicates is 0.9903, outperforming existing techniques.

Images: 100 Gb per lane - 800 Gb per flowcell IPAR Output: 10.4 Gb/lane - 83.2 Gb/flowcell Intensity files: 8.9 Gb/lane - 71.2 Gb/flowcell Basecalls: 22 Gb/lane - 176 Gb/flowcell SRF: 7.53 Gb/lane - 60.42 Gb/flowcell Filtered Sequences: 1.6 Gb/lane - 12.8 Gb/flowcell Alignment exports: 1.23 Gb/lane - 14.76 Gb/flowcell Error reports: 6.47 Gb/lane - 51.76 Gb/flowcell Minimal Dataset: 8.76 Gb/lane - 70 Gb/flowcell Everything without images: 66.89 Gb/lane - 535 Gb/flowcell Everything including images: 166.7 Gb/lane - 1.3 Tb/flowcell

mage by Werner Van Belle