Deep Sequencing

Werner Van Belle[1], Ina Nissen[1], Michael Stadler[2] and Christian Beisel[1]

(1) Laboratory for Quantitative Genomics, Department Biosystems Science and Engineering - ETH Zurich
(2) Friedrich Miescher Institute for Biomedical Research

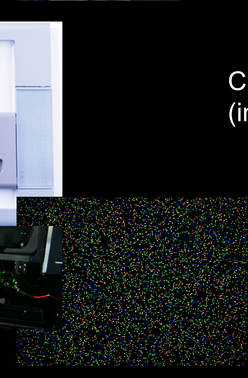
D-BSSE
Department of Biosystems Science and Engineering

Abstract

In parallel to the human genome sequencing initiative several new technologies have emerged that allow sequencing at unprecedented throughput and low costs. These approaches are generally referred to as “deep sequencing”. They enable researchers to not only re-sequence genomes and thus to identify genome variations but also to quantify the abundance of experimentally enriched fractions of the genome. The very large numbers of short individual sequence reads produced by the Illumina Genome Analyzer (currently approx. 50 million reads per instrument run) are well suited to make direct quantitative measurements of the sequence content of a DNA sample. By determining a short sequence read from each of many randomly selected molecules from the sample and then mapping each sequence read onto the reference genome, the identity of each starting molecule is learned, and its frequency in the sample can be calculated. Desired levels of sensitivity and statistical certainty, needed to detect rare molecular species, can be achieved by adjusting the total number of sequence reads. Sequence census assays do not require knowing in advance that a sequence is of interest as a promoter, enhancer or RNA-coding domain, as most current microarray designs do. The combination of chromatin immunoprecipitation assays with the subsequent quantitative analysis of the enriched DNA sample by deep sequencing (ChIP-seq) has been proven to be of great value for whole mammalian genome approaches in several high-profile studies published over the last year. At D-BSSE we have established a deep sequencing unit based on Illumina sequencing technology located in the new “Laboratory for Quantitative Genomics”. This poster gives an overview of the sequencing technology and the data analysis pipeline. Furthermore it provides insights into the quality of our functional genomics data recently generated by ChIP-seq and RNA-seq.

Workflow and Applications

DNA
↓
Sample processing (DNA library)
↓
Cluster generation
↓
Sequencing by synthesis
↓
Data analysis



Chromatin profiling (ChIP-seq)

Expression profiling (RNA-seq, miRNA)

Genome resequencing

De novo Sequencing

Chromosomal rearrangements (insertions, deletions, translocations)

SNP detection

Copy number variations

DNA methylation

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).

Each lane has 100 imaging positions, called tiles, spaced in two columns

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

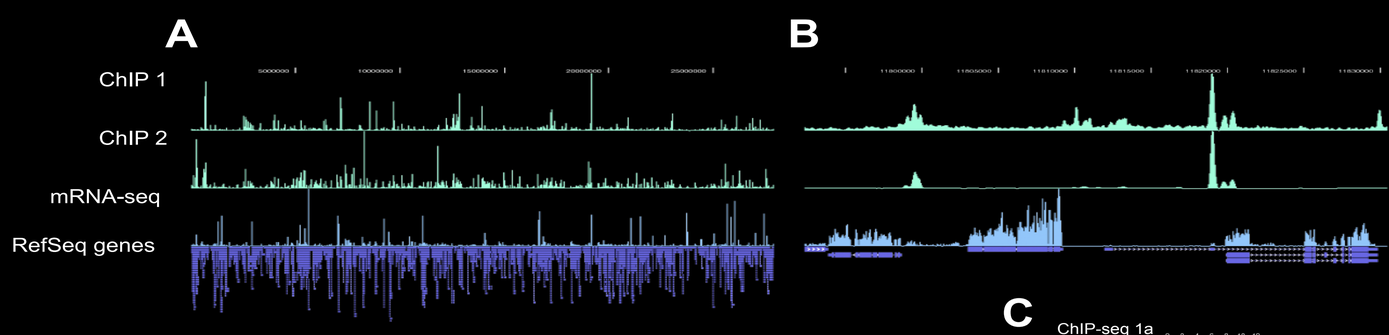
4 tile images. Top left: normal size. Top right: 2x. Bottom left: 50x. Bottom right 100x zoom. Each spot is a short fragment (cluster)

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

Short Fragment Alignment

The sequenced short fragments can be aligned to a reference genome (yellow). The alignment program “Eland” allows for at most two mismatches per fragment. “PhageAlign” finds the best match for each fragment.

Chromatin Profiling



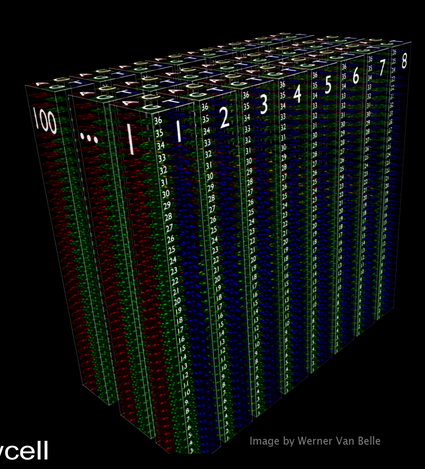
Chromatin profiling of two epigenetic regulators of the Polycomb and Trithorax group. A, Protein distribution and mRNA detection across 30 million bases of the *Drosophila* genome. B, zoom in (35k). C, technical reproducibility of ChIP sequencing. One processed ChIP sample was sequenced three times.

Expression Profiling

	Exon Mass	Length	Intron Mass	Length	Expression Total	0.961517222 Exon	CG	Description
3	3309228	814	304795	183	3624.897693	4065.390663	CG9282	60S ribosomal protein L24. [Source:Uniprot/SWISSPROT;Acc:Q9VJY6]
4	2776390	798	748760	965	1999.517867	3479.185464	CG7808	40S ribosomal protein S8. [Source:Uniprot/SWISSPROT;Acc:Q8MLY8]
5	3113478	1011	280139	448	2325.988348	3079.602374	CG11522	Ribosomal protein L6 CG11522-PB, isoform B [Source:RefSeq_peptide;Acc:NP_651876]
6	2726923	908	162726	265	2463.468883	3003.219163	CG6779	40S ribosomal protein S3. [Source:Uniprot/SWISSPROT;Acc:Q06559]
7	4340876	1458	499825	672	2272.629577	2977.281207	CG5502	60S ribosomal protein L4 (L1). [Source:Uniprot/SWISSPROT;Acc:P09180]
8	7535847	2676	44729	1297	1908.023156	2816.086323	CG4264	Heat shock 70 kDa protein cognate 4 (Heat shock 70 kDa protein 8E). [Source:Uniprot/SWISSPROT;Acc:P11147]
9	5002990	1832	93038	491	2193.727077	2730.889738	CG1913	Tubulin alpha-1 chain. [Source:Uniprot/SWISSPROT;Acc:P06603]
10	1953664	723	368017	260	2361.832146	2702.163209	CG8495	40S ribosomal protein S29. [Source:Uniprot/SWISSPROT;Acc:Q9VH69]
11	1995472	739	357882	755	1575.203481	2700.232747	CG1475	60S ribosomal protein L13A. [Source:Uniprot/SWISSPROT;Acc:Q9VNE9]
12	1540676	576	369819	310	2156.314898	2674.784722	CG15697	Ribosomal protein S30 CG15697-PB, isoform B [Source:RefSeq_peptide;Acc:NP_732566]
13	2251277	844	94166	510	1732.232644	2667.38981	CG4759	Ribosomal protein L27 CG4759-PA [Source:RefSeq_peptide;Acc:NP_651417]
14	1930520	736	306020	360	2040.638686	2622.98913	CG6684	40S ribosomal protein S25. [Source:Uniprot/SWISSPROT;Acc:P48588]
15	3039445	1317	393818	1186	1371.659209	2307.854973	CG1883	40S ribosomal protein S7. [Source:Uniprot/SWISSPROT;Acc:Q9VA91]
16	892566	403	221425	60	2406.028078	2214.80397	CG13874	General odorant-binding protein 56h precursor. [Source:Uniprot/SWISSPROT;Acc:Q9VBY9]
17	2181894	1031	133	1	2114.367248	2116.28904	CG4800	Translationally-controlled tumor protein homolog (TCTP). [Source:Uniprot/SWISSPROT;Acc:Q9VGS2]
18	1268885	675	354140	856	1060.107773	1879.82963	CG2099	Ribosomal protein L35A CG2099-PA [Source:RefSeq_peptide;Acc:NP_649539]
19	2227098	1204	41	1	1848.248133	1849.749169	CG9010	CG9010-PA [Source:RefSeq_peptide;Acc:NP_611172]
20	1542457	856	302438	157	1821.219151	1801.935748	CG7939	60S ribosomal protein L32 (Ribosomal protein 49). [Source:Uniprot/SWISSPROT;Acc:P04359]
21	974040	562	494782	616	1246.877759	1733.16726	CG15693	40S ribosomal protein S20. [Source:Uniprot/SWISSPROT;Acc:P55828]
22	2965406	1752	131243	196	1589.655544	1692.583333	CG11901	Elongation factor 1-gamma (EF-1-gamma) (eEF-1B gamma). [Source:Uniprot/SWISSPROT;Acc:Q9NJH0]
23	1692571	1016	98866	103	1600.92672	1665.916339	CG2168	40S ribosomal protein S3a (C3 protein). [Source:Uniprot/SWISSPROT;Acc:P55830]
24	2088834	1261	585261	350	1659.897579	1656.490087	CG8592	Protein stand still. [Source:Uniprot/SWISSPROT;Acc:P92189]
25	1772259	1086	1581858	327	2373.755839	1631.914365	CG6048	CG6048-PA [Source:RefSeq_peptide;Acc:NP_572282]
26	1149387	715	345676	834	965.1794706	1607.534266	CG10423	Ribosomal protein S27 CG10423-PA [Source:RefSeq_peptide;Acc:NP_651359]
27	1365852	881	108401	65	1558.406977	1550.342792	CG9354	RpL34b CG9354-PA, isoform A [Source:RefSeq_peptide;Acc:NP_649887]
28	3280145	2441	526809	839	1160.656707	1343.770995	CG4863	60S ribosomal protein L3. [Source:Uniprot/SWIS
29	2528588	1914	276867	118	1380.637303	1321.101358	CG13436	CG13436-PA [Source:RefSeq_peptide;Acc:NP_61
30	1839192	1463	206417	488	1048.492568	1257.137389	CG9429	Calreticulin precursor (CRP55) (Calregulin) (HAC
31	1597329	1478	226519	4658	297.2372881	1080.736806	CG31196	14-3-3 protein epsilon (Suppressor of Ras1 3-9).
32	1492361	1434	7	1	1039.9777	1040.698047	CG32564	CG32564-PA [Source:RefSeq_peptide;Acc:NP_72
33	1172022	1167	106085	1618	458.9253142	1004.303342	CG6378	BM-40-SPARC CG6378-PA [Source:RefSeq_pepti
34	672760	681	67429	181	858.687935	987.9001468	CG2210	Nucleoside diphosphate kinase (EC 2.7.4.6) (ND
35	863989	897	795	1	963.0111359	963.1984392	CG18180	CG18180-PA [Source:RefSeq_peptide;Acc:NP_64
36	838186	906	228196	1072	539.1213347	925.1501104	CG5499	Histone H2A variant. [Source:Uniprot/SWISSPRC
37	3721111	4267	514557	223	943.355902	872.0671604	CG12750	CG12750-PA [Source:RefSeq_peptide;Acc:NP_6
38	3800526	4407	9414	358	799.567681	862.3839346	CG1531	CG1531-PB [Source:RefSeq_peptide;Acc:NP_572
39	986215	1210	7983	509	578.3583479	815.053719	CG10045	Glutathione S-transferase 1-1 (EC 2.5.1.18) (GS
40	5687806	7059	238955	636	770.2093567	805.7523739	CG14967	CG14967-PA [Source:RefSeq_peptide;Acc:NP_64
41	4995902	6262	88822	3073	544.6945903	797.81252		
42	1827849	2353	423710	896	693.0006156	776.8164046	CG3002	Gga CG3002-PB [Source:RefSeq_peptide;Acc:NP
43	505664	656	167769	217	771.4009164	770.8292683	CG14115	CG14115-PA [Source:RefSeq_peptide;Acc:NP_64
44	1208685	1600	1217459	979	940.7305157	755.428125	CG17746	CG17746-PB, isoform B [Source:RefSeq_peptide
45	520710	694	33985	124	678.1112469	750.3025937		
46	536165	720	139019	333	641.2003799	744.6736111		
47	648988	873	819	1	743.48627	743.3997709	CG18179	CG18179-PA [Source:RefSeq_peptide;Acc:NP_64
48	1537866	2084	57154	114	717.8885507	739.6876291	CG33251	CG33251-PA [Source:RefSeq_peptide;Acc:NP_72

By counting the number of sequenced bases matched at a certain genome position, the expression level of each gene is calculated. The correlation between two technical replicates is 0.99 and the dynamic range comprises four to five orders of magnitude.

Data Delivery



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