Linking FANTOM5 CAGE peaks to annotations with CAGEscan

Nicolas Bertin[1,2,3], Mickaël Mendez[1,2,4], Akira Hasegawa[1,2], Marina Lizio[1,2], Imad Abugessaisa[1,2], Jessica Severin[1,2], Mizuho Sakai-Ohno[1,2], Timo Lassmann[1,2,5], Takeya Kasusawa[1], Hideya Kawaji[1,2,6], Yoshihide Hayashizaki[1,2,6], Alistair R. R. Forrest[1,2,7], Piero Carninci[1,2], Charles Plessy[1,2*]

April 11, 2017

1. RIKEN Center for Life Science Technologies, Division of Genomics Technologies, Japan; 2. RIKEN Omics Science Center, Japan; 3. Present address: Human Longevity Singapore Pte. Ltd., Singapore; 4. Present address: Department of Computer Science, University of Toronto, Canada; 5. Present address: Telethon Kids Institute, The University of Western Australia, Australia; 6. RIKEN Preventive Medicine and Diagnosis Innovation Program, Japan; 7. Present address: Harry Perkins Institute of Medical Research, QEII Medical Centre and Centre for Medical Research, the University of Western Australia, Australia; *Corresponding author(s): Charles Plessy (plessy@riken.jp).

Abstract

The FANTOM5 expression atlas is a quantitative measurement of the activity of nearly 200,000 promoter regions across nearly 2,000 different human primary cells, tissue types and cell lines. Generation of this atlas was made possible by the use of CAGE, an experimental approach to localise transcription start sites at single-nucleotide resolution by sequencing the 5′ ends of capped RNAs after their conversion to cDNAs. While 50% of CAGE-defined promoter regions could be confidently associated to adjacent transcriptional units, nearly 100,000 promoter regions remained gene-orphan. To address this, we used the CAGEscan method, in which random-primed 5′-cDNAs are paired-end sequenced. Pairs starting in the same region are assembled in transcript models called CAGEscan clusters. Here, we present the production and quality control of CAGEscan libraries from 56 FANTOM5 RNA sources, which enhances the FANTOM5 expression atlas by providing experimental evidence associating core promoter regions with their cognate transcripts.

Background & Summary

CAGE (Cap Analysis Gene Expression, [1]) is the method of choice for studying gene regulation through quantitative analysis of transcription start sites (TSS, sequence ontology term 0000315) [2]. By sequencing the 5′ end of cDNA-converted capped RNAs, CAGE enables the identification of core promoter regions and 5′ end transcriptional activity. Large scale application of CAGE by

the FANTOM consortium to nearly 2,000 human RNA sources including primary cells, whole-tissue extracts and cell lines [3, 4] identified nearly 200,000 core promoter regions active within the human genome [5].

Although CAGE enables the location of TSS at a single nucleotide resolution, the determination of their connection to downstream known gene structures or to independent novel RNAs is limited to positional computational inference and low-throughput gene-by-gene experimental validations. Half (101,893/201,802) of the FANTOM5's active core promoter regions did not co-localize within a reasonable distance with 5′ termini of annotated gene models. To experimentally associate these orphan core promoter regions to transcriptional units, we employed *CAGEscan* [6], an approach in which paired-end sequencing of the 5′ end of cDNA-converted capped RNAs with their cognate randomly priming sites enables the unequivocal association of individual TSS to transcripts exons. In a previous project, focused on analysing the translatome of Purkinje neurons in rat [7], the CAGEscan approach annotated 43 % of the core promoters active in rat's Purkinje neurons that we detected but had no by direct overlap with Ensembl transcripts.

Here, we selected 56 RNA sources which upon FANTOM5 CAGE profiling revealed the greatest levels of transcriptome diversity and prepared individual CAGEscan libraries, with 6 of these 56 RNA sources prepared in duplicate (see Table 1). Using the FANTOM5 core promoter atlas as seed, we clustered the CAGEscan paired-end reads in a collection of 112,315 models called *CAGEscan clusters*. To de-orphanise FANTOM5 promoters, we intersected the CAGEscan clusters with GENCODE 18 gene models. Of the 85 % that intersected, 33,632 clusters had no annotation in FANTOM5, thus revealing novel and alternative promoters to known genes. We made these data available along with the FANTOM5 CAGE atlas data, as well as ready for manual inspection and analysis via the ZENBU genome browser [8] (see Figure 1 and Data Citation 1).

	Source.Name	Description	Unextracted	Artefacts	rDNA	Non_aligned	Non_proper	Duplicates	Promoter	Exon	Non_annotated
NCig10013	10002-101A5	SABiosciences XpressRef Human Universal Total RNA, pool1	865232	53578	8620490	303381	336125	914464	352035	557012	978157
NCig10014	10012-101C3	brain, adult, pool1	789460	56053	3579617	156712	499657	1967826	479841	1007678	1505064
NCig10015	10016-101C7	heart, adult, pool1	657065	67493	12680315	189587	360818	1791722	341589	679820	303502
NCig10016	10026-101D8	testis, adult, pool1	735402	56663	8828467	229250	357108	761921	321059	575393	913618
NCig10017	10030-101E3	retina, adult, pool1	983120	49209	2016040	91931	396226	1395594	341442	574800	1590536
NCig10018	11210-116A4	Smooth Muscle Cells - Aortic, donor0	636805	76008	14079255	126255	163413	505347	152188	219216	111093
NCig10019	12176-128I7	Whole blood (ribopure), donor090325, donation1	745633	59630	5626776	168819	557490	1749239	502772	701914	1608998
NCig10020	10019-101D1	lung, adult, pool1	853743	90406	8943567	141998	347985	819745	309787	599126	1087789
NCig10021	10022-101D4	prostate, adult, pool1	746116	64413	6095769	167516	486536	1152553	366192	716651	1338368
NCig10022	10025-101D7	spleen, adult, pool1	852309	62976	3130575	143638	431102	1038248	402237	785577	1493319
NCig10023	10150-102I6	medial frontal gyrus, adult, donor10252	960285	25800	467916	81388	389764	534385	224651	506845	1320993
NCig10024	10151-102I7	amygdala, adult, donor10252	858652	33503	1112179	112044	450492	801134	321988	702445	1921642
NCig10025	10153-102I9	hippocampus, adult, donor10252	892716	32233	861081	80184	360888	643485	241503	530452	1425771
NCig10026	10154-103A1	thalamus, adult, donor10252	817751	35872	2262386	95945	505911	1296220	377511	716479	2043883
NCig10027	10155-103A2	medulla oblongata, adult, donor10252	1397247	78367	2366286	132873	610522	1541331	446644	906737	2519780
NCig10028	10157-103A4	parietal lobe, adult, donor10252	936084	71291	1471269	153886	694447	1387954	463355	1096767	3181090
NCig10029	10158-103A5	substantia nigra, adult, donor10252	1078146	59251	2360698	85562	425712	1188939	344322	602764	1511269
NCig10030	10159-103A6	spinal cord, adult, donor10252	888981	79200	2801018	116324	594991	1353570	442320	903764	2471015
NCig10031	10160-103A7	pineal gland, adult, donor10252	1011960	65055	1343792	103948	545004	944389	348323	744959	2470004
NCig10032	10161-103A8	globus pallidus, adult, donor10252	821077	80387	4015936	130251	673588	1632249	469685	916587	2749739
NCig10033	10162-103A9	pituitary gland, adult, donor10252	970964	49755	1932932	124341	563707	1591606	455105	847858	2093988
NCig10034	10163-103B1	occipital cortex, adult, donor10252	905254	44694	1193623	136650	708731	1223230	432869	1030595	3731863
NCig10035	10164-103B2	caudate nucleus, adult, donor10252	1102408	38186	869711	109813	476042	1310808	346046	754780	1809163
NCig10036	10165-103B3	locus coeruleus, adult, donor10252	1045453	49711	1251961	97962	454490	1173365	330395	729525	1620164
NCig10037	10166-103B4	cerebellum, adult, donor10252	1095992	54415	368370	62650	519173	650125	254418	492016	2527876
NCig10038	11207-116A1	Endothelial Cells - Aortic, donor0	718897	98322	12128937	142908	291064	820844	298247	455971	306374
NCig10039	11222-116B7	Fibroblast - Gingival, donor4 (GFH2)	885111	167833	2081881	108043	284519	1080129	346774	547431	363853
NCig10040	11224-116B9	CD14+ Monocytes, donor1	651017	101461	4297440	152438	540035	1268085	510000	645877	3065822
NCig10041	11229-116C5	CD14+ monocyte derived endothelial progenitor cells, donor1	1032309	242145	3950479	190791	539087	1666613	561795	835546	1756556
NCig10042	11245-116E3	Fibroblast - Aortic Adventitial, donor1	735498	828670	3376827	198604	517858	2135913	655705	710317	384044
NCig10043	11246-116E4	Intestinal epithelial cells (polarized), donor1	919980	392820	1056095	120525	433407	2003503	513395	536761	705255
NCig10044	11247-116E5	Mesothelial Cells, donor1	870202	443481	1547197	127726	418103	2291855	516801	516012	419344
NCig10045	11248-116E6	Anulus Pulposus Cell, donor1	673123	467283	4733836	191255	418230	1674689	474236	571373	125453
NCig10046	11249-116E7	Pancreatic stromal cells, donor1	895678	266841	1598919	129096	447633	1839323	563482	606168	570720
NCig10047	11256-116F5	Small Airway Epithelial Cells, donor1	762215	197286	3113793	175723	506591	2330196	629638	801987	416965
NCig10048	11273-116H4	Mammary Epithelial Cell, donor1	890533	198834	4561811	208742	497865	2025351	497139	721321	417785
NCig10049	11278-116H9	Placental Epithelial Cells, donor1	523668	434079	7019440	196493	358154	1908353	304347	296384	171089
NCig10050	11282-116I4	Skeletal muscle cells differentiated into Myotubes - multinucleated, donor1	825574	278816	3852864	174167	447392	2002086	521214	534883	274710
NCig10051	11468-119C1	Preadipocyte - omental, donor1	863018	244070	1743473	94850	257299	790493	304494	524536	287355
NCig10052	11487-119E2	Mast cell - stimulated, donor1	1047459	53428	390687	86272	312008	1219897	244001	294922	739794
NCig10053	10411-106B6	renal cell carcinoma cell line:OS-RC-2	774316	209703	2297666	117997	421779	1058325	387862	580214	1057849
NCig10054	10412-106B7	malignant trichilemmal cyst cell line:DJM-1	728347	139130	3031554	164712	630434	2045672	648552	895267	1574617
NCig10055	10414-106B9	maxillary sinus tumor cell line:HSQ-89	857517	135611	2250778	123989	579817	1951209	498578	697019	2030545
NCig10056	10431-106D8	epidermoid carcinoma cell line:Ca Ski	1071422	146593	982637	87543	343508	859004	378966	452570	752743
NCig10057	10436-106E4	signet ring carcinoma cell line:Kato III	840579	145687	3244444	137763	512628	1657263	503623	614540	1037414
NCig10058	10442-106F1	schwannoma cell line:HS-PSS	941029	176159	1799562	134668	519866	1659980	589180	733263	1160911
NCig10059	10444-106F3	glioblastoma cell line:A172	861701	175094	2804921	186736	495954	1209931	520670	712755	1298299
NCig10060	10454-106G4	chronic myelogenous leukemia cell line:K562	1045797	109272	645627	70593	363272	675740	342295	400380	1103605
NCig10061	10464-106H5	acute lymphoblastic leukemia (T-ALL) cell line:Jurkat	869111	131089	2562674	178216	687478	1748129	774916	819425	1573041
NCig10062	10508-107D4	neuroblastoma cell line:CHP-134, tech_rep1	962974	148947	278618	57421	405741	662738	258098	391938	1456216
NCig10063	10552-107I3	cervical cancer cell line:D98-AH2, tech_rep1	1005845	156179	421514	70319	310350	1186016	271445	368888	516869
NCig10064	10558-107I9	osteosarcoma cell line:HS-Os-1, tech_rep1	983856	182894	548737	80879	357116	711651	286493	395130	827321
NCig10065	10410-106B5	extraskeletal myxoid chondrosarcoma cell line:H-EMC-SS, tech_rep1	928677	138036	393526	64950	343707	582912	220600	400189	892753
NCig10066	10441-106E9	synovial sarcoma cell line:HS-SY-II, tech_rep1	844408	197018	574814	57821	348974	523331	235006	375904	882555
NCig10067	10474-106I6	myeloma cell line:PCM6, tech-rep1	810459	186856	755594	71301	358371	807453	278280	416507	897364
NCig10068	10424-106D1	splenic lymphoma with villous lymphocytes cell line:SLVL	852283	163002	455663	80461	376376	969585	280831	377707	903091
NCig10126	10508-107D4	neuroblastoma cell line:CHP-134, tech_rep2	995795	48625	550450	63938	396327	701319	298112	438554	1766026
NCig10127	10552-107I3	cervical cancer cell line:D98-AH2, tech_rep2	1015542	60740	646950	64609	304041	930823	243837	338193	492654
NCig10128	10558-107I9	osteosarcoma cell line:HS-Os-1, tech_rep2	968282	75235	865891	76828	336463	737523	296891	409283	915232
NCig10129	10410-106B5	extraskeletal myxoid chondrosarcoma cell line:H-EMC-SS, tech_rep2	822752	40614	436600	112425	377956	276992	152854	276872	621505
NCig10130	10441-106E9	synovial sarcoma cell line:HS-SY-II, tech_rep2	726773	64905	633633	81424	473478	240861	160678	250046	600963
NCig10131	10474-106I6	myeloma cell line:PCM6, tech-rep2	720124	60988	898043	78483	322369	566406	231781	346042	761108

Table 1: Summary of the libraries prepared. The RNA identifier (Source.Name) can be searched in the FANTOM5 SSTAR database [9, 10]. The RNA samples are also described in the SDRF files distributed alongside the FASTQ sequences and alignments, as well as the raw alignment statistics.

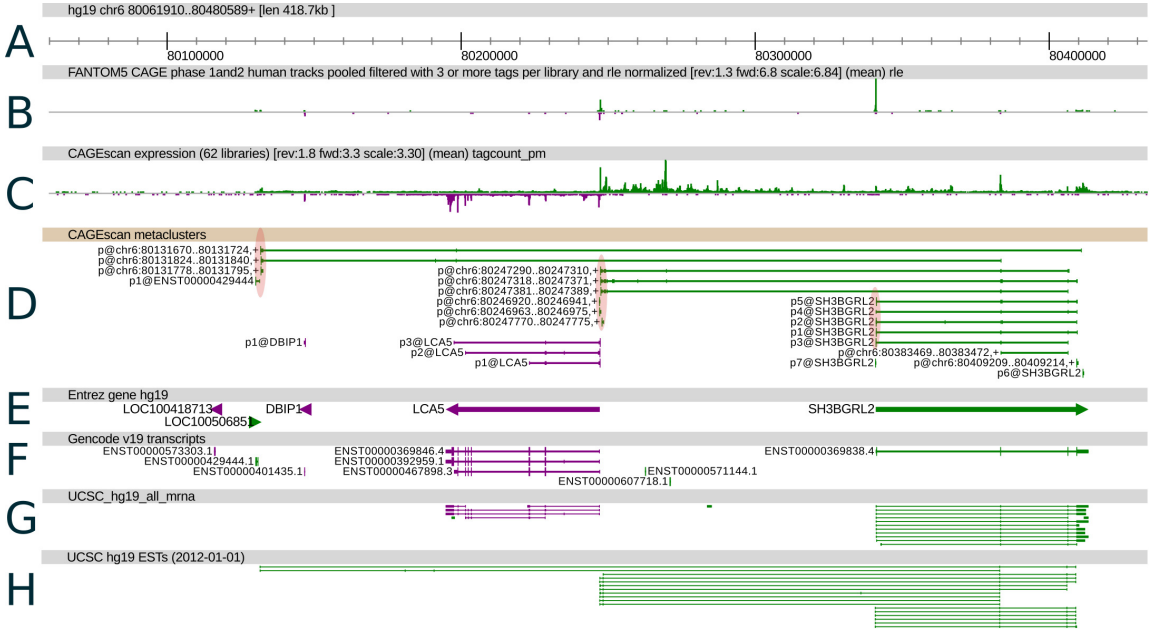

Figure 1: CAGEscan clusters revealing new promoters for the SH3BGRL2 gene. Features on the plus and minus strand are displayed in green and purple respectively. Promoter regions of interest are highlighted with ellipses in track D. A: Genomic coordinates. B: FANTOM5 CAGE signal as a quantitative histogram. C: CAGEscan CAGE signal. D: CAGEscan meta-clusters, combining pairs for all libraries. The name of the seed CAGE peak is indicated on the left of each cluster. E: NCBI Gene bodies. F: GENCODE 19 annotations. G: GenBank mRNA sequences. H: EST sequences supporting the CAGEscan clusters.

Methods

All human samples used in the project were either exempted material (available in public collections or commercially available), or provided under informed consent. All non-exempt material is covered under RIKEN Yokohama Ethics applications (H17-34 and H21-14). The CAGEscan libraries were prepared as described earlier [11]. In brief, 500 ng of RNA were reverse-transcribed in presence of random primers and template-switching oligonucleotides, amplified by PCR and sequenced paired-end (2×36 nt) on Illumina GAIIx sequencers, one sample per lane. The barcode sequence `GCTATA`, present in every sample, acted as the spacer that we introduced in [12] to decrease the amount of strand-invasion artifacts. The paired-end sequences were then processed with the MOIRAI workflow system [13], with a template implementing the workflow OP-WORKFLOW-CAGEscan-FANTOM5-v1.0, described below and in Figure 2.

For each pair, the first (CAGE) and second (CAGEscan) reads in FASTQ format were demultiplexed. The first 9 bases of the CAGE reads were trimmed as they contain the sample barcode and the template-switching linker. CAGEscan paired-end reads that did not contain the exact barcode and linker sequences were discarded. The first 6 bases of the CAGEscan reads were trimmed,

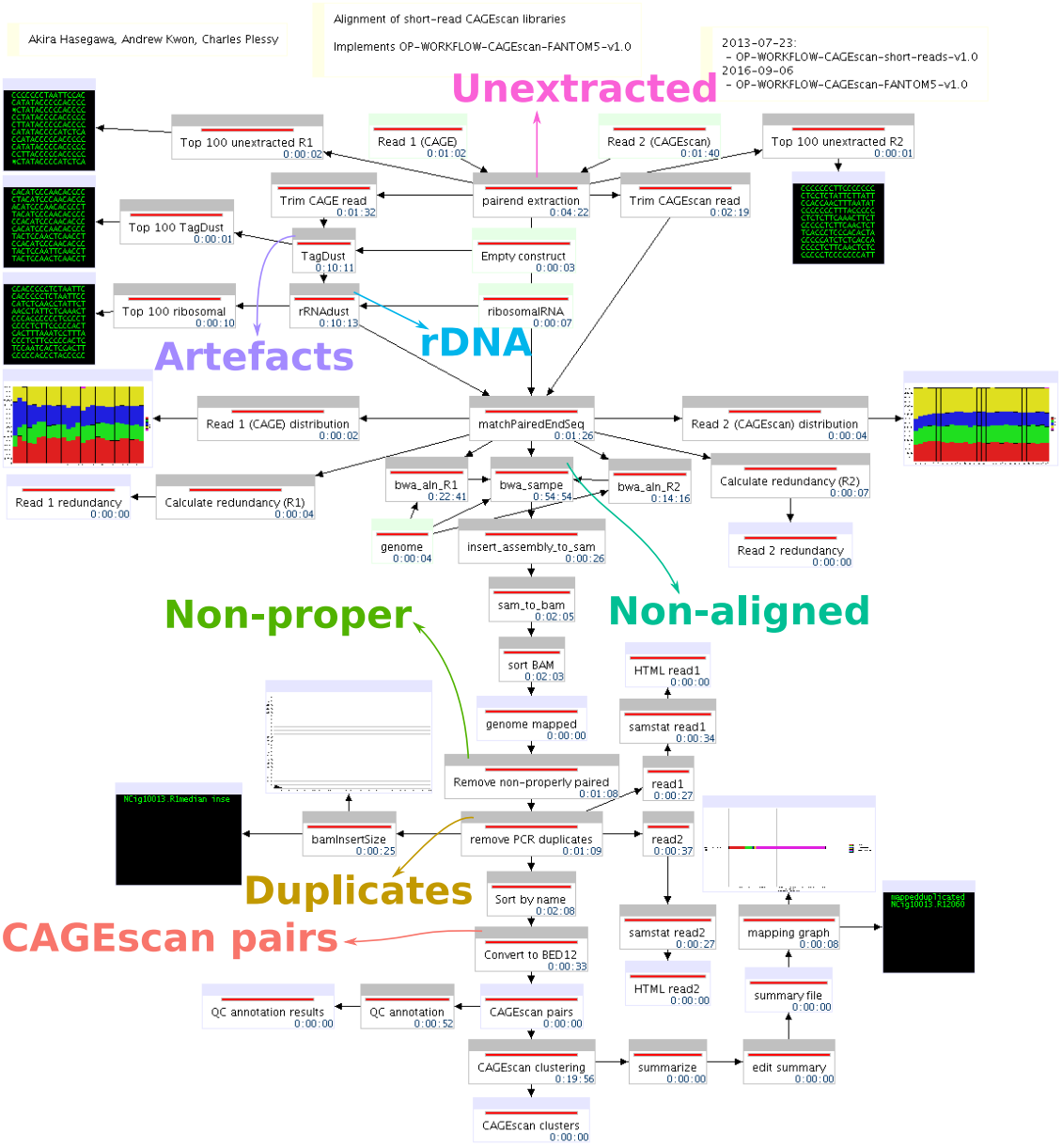

Figure 2: Processing pipeline. The diagram made of boxes connected by black arrows displays the MOIRAI workflow completed for one (NCig10013) of the 62 CAGEscan libraries. The colored text and arrows overlayed on the diagram represents the points where the main alignment statistics are calculated to summarize the number of read pairs passing all the filters (CAGEscan pairs) or discarded at each step of the processing pipeline (Unextracted, rDNA, Artifacts, Non-aligned, Non-proper, Duplicates).

because they originate from the random primers and not the cDNAs, and therefore are prone to errors caused by mismatches during the hybridization to the RNAs, that are well tolerated by the reverse-transcriptase [14].

The CAGE and CAGEscan reads were then filtered independently with the TagDust program version 1.13 [15], using the sequences of empty constructs and primers as artifact library. They were then compared to reference sequences of ribosomal genes (GenBank: U13369.1) using the `rRNAdust` program version 1.03. Reads whose mates were discarded by these two filters were then removed.

FASTQ formatted cleaned paired-end reads were then aligned on the human genome version `hg19` with BWA version 0.7.15 [16] using standard parameters, except that the maximum insert length (`-a`) was set to 2 Mbp to allow pairs to map on different exons, and that insert size detection was disabled (`-A`). Extra header records (for SQ: AS and for RG: CN, ID, LB, PU, SM, and PL) were added to ease processing and tracking. The resulting BWA SAM formatted alignments were then converted to BAM format, and unmapped as well as non-properly paired CAGE reads were discarded (flag `0x42`). The resulting "CAGEscan pairs" provide individual experimental information on the association of a single-nucleotide-resolution TSS with the body of a gene product.

The CAGEscan pairs were then converted to BED12 format using the program `pairedBamToBed12` version 1.2, in which the score field is the sum of the mapping qualities of each read of the pair. They were then assembled into CAGEscan clusters using the `CAGEscan-Clustering` script version 1.2 and the Phase 1 + 2 FANTOM5 DPI CAGE peaks as seeds. The CAGEscan-Clustering script also takes advantage of the BED12 format, reporting the number of CAGEscan paired-end reads used to assemble each cluster via the score field and the name and position of the seeding CAGE peak via the name, thickStart and thickEnd fields respectively. Finally, the CAGEscan clusters from all libraries were then combined into a single global assembly of "meta-clusters" using the same program and output in BED12 files where the score indicates the number of libraries contributing data to each meta-cluster.

Code availability

The MOIRAI workflow template used to process the libraries is available as a supplemental XML file on Figshare (DOI: 10.6084/m9.figshare.4792666). MOIRAI enabled the design of a complete data processing pipeline based on the following softwares: FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), TagDust 1.13 [15], rRNAdust 1.03 (http://fantom.gsc.riken.jp/5/sstar/Protocols:rRNAdust) (note that for new projects, we recommend TagDust 2 instead of TagDust 1 and rRNAdust), BWA 0.7.15-r1140 ([16]), SAMtools 0.1.19-44428cd ([17]), pairedBAMtoBED12 1.2 (https://github.com/Population-Transcriptomics/pairedBamToBed12, DOI: 10.6084/m9.figshare.4792672), CAGEscan-Clustering.pl 1.2 (https://github.com/nicolas-bertin/CAGEscan-Clustering, DOI: 10.6084/m9.figshare.4792675) and promexinstats.sh for the annotation (see supplemental material). The software above and standard Unix tools are sufficient to re-implement the pipeline in a different workflow system.

Data Records

Each CAGEscan library is described with a Sample and Data Relationship Format (SDRF) record, together with the rest of the FANTOM5 data ([9]). For each library, raw sequences in FASTQ format, alignment data in BAM format (including unmapped reads), CAGEscan pairs in BED12 format, CAGEscan clusters in BED12 format and alignment statistics in plain text tabulation-

delimited triples (subject, predicate, object), are available in the FANTOM5 data repository. The raw sequences have also been deposited to DDBJ DRA (Data Citation 2).

Technical Validation

We derived individual library alignment statistics from the MOIRAI data processing pipeline (see Table 1 and Figures 2 and 3A). The statistics count the number of reads discarded at key steps of the processing. "Unextracted" are pairs where the linker was not found, "Artefacts" are pairs that matched the artifact library, or had a low complexity, "rDNA" are pairs that matched the reference rDNA locus (including rRNAs and their spacer regions), "Non-aligned" are pairs where one or both mates were not aligned to the genome, and "Non-proper" are pairs where the mates were not aligned in head-to-head orientation within 2 Mbp. "Duplicates" are the pairs removed during the deduplication step. That is, when there are n pairs with identical coordinates, 1 is kept and $n-1$ are discarded as "Duplicates". These statistics show that the amount of PCR duplicates was not larger than the number of CAGEscan pairs, suggesting that the libraries prepared in this study have not been fully exhausted by sequencing.

The library alignment statistics, as well as statistics describing the distribution of CAGEscan TSSs on GENCODE 19 annotations (Figure 3B), also suggest that the biological nature of the samples (cancer cell lines, primary cells, tissue samples and brain tissue) strongly influenced the performance of the CAGEscan protocol used in this study. Albeit displaying the best performance in terms of alignment (largest fraction of CAGEscan pairs), brain tissue derived samples had the lowest rate of known promoters overlapping start sites, hinting at a much greater diversity of alternative promoters usage in human brain. However, since, in this study, all brain tissue derived samples were taken from a single donor, this observation may result from technical batch effect rather than being a general feature of the nature of human brain transcriptome.

Usage Notes

We have seeded the CAGEscan clustering with FANTOM5 CAGE-defined core promoter regions, however alternative seeding strategies could be envisioned. The $5'$ ends of the CAGEscan pairs themselves could be clustered by peak calling and used as a seed, which is the default mode of operation of the `pairedBamToBed12` tool. Foregoing the discovery of alternative promoters, CAGEscan clusters could also be seeded using promoter regions defined by GENCODE models. To discover potential enhancer-associated non-coding RNAs, region corresponding to FANTOM5 enhancers [18] could also be used.

We used a simple alignment strategy that did not take splicing into account. Thus, pairs overlapping splice junctions could not be mapped and CAGEscan clusters lack coverage at the beginning and end of each exon, but this only mildly impacts the main purpose of the method. In addition, since the CAGEscan pairs are anchored at the $5'$ end of the transcripts, splice junctions occurring close to the TSS may render some whole loci unmappable. Indeed, transcripts databases such as GENCODE reveal splice junctions very near to the TSS. Trimming the CAGE reads to 20 nt rescued some loci, but other loci were lost due to the decrease of alignment stringency (data not shown). Thus, the development of a spliced alignment workflow would increase the accuracy of our method.

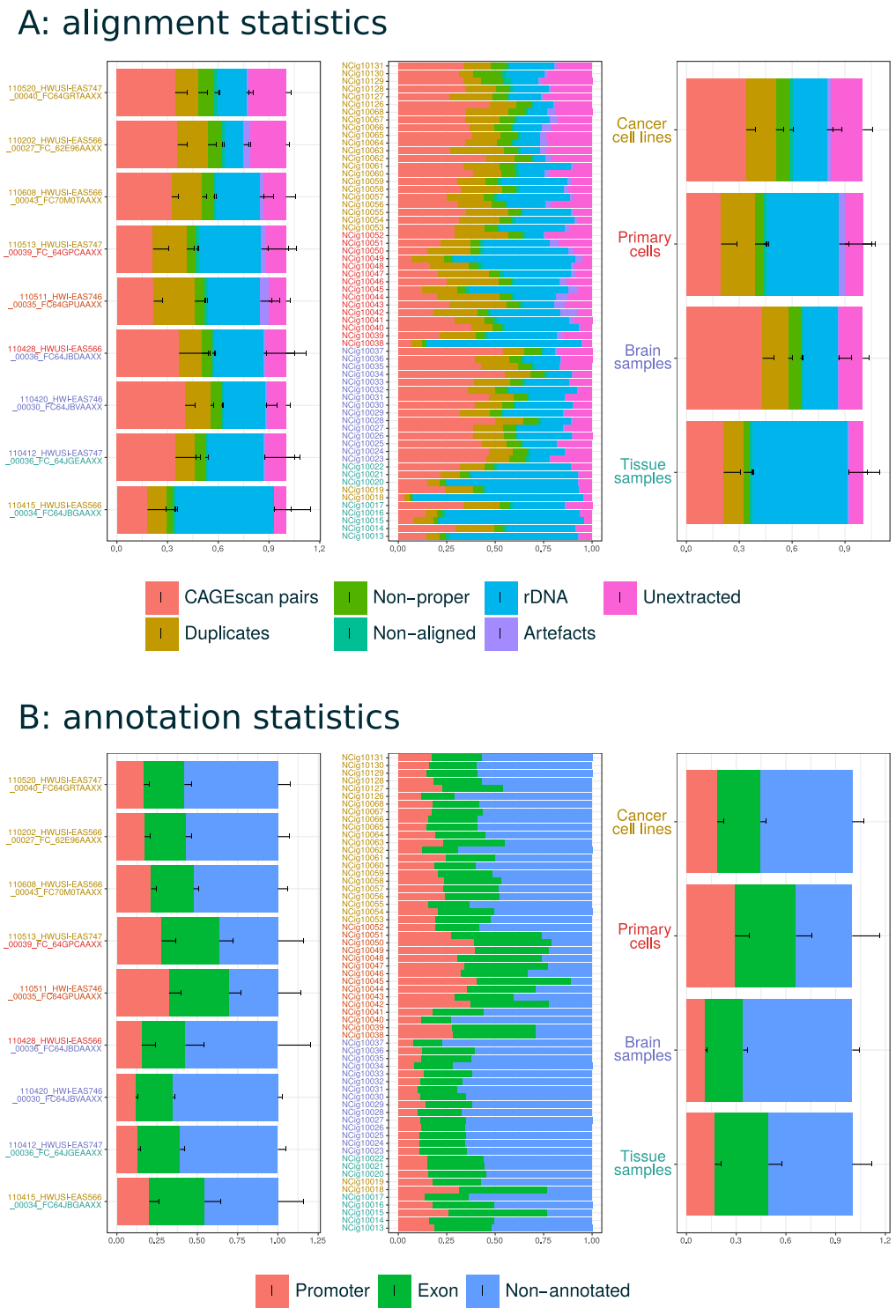

Figure 3: Quality control statistics. A: Fraction of pairs passing all filters (CAGEscan pairs) or discarded at key steps of the processing pipeline (see Figure 2). The central block of stack bars represents each library individually. The left block aggregates them by sequencing batch, named by the sequencing run identifier. The right block aggregates the libraries by sample type. Each sample type is represented by one color, that is also used to color the library identifiers and the sequence identifiers in the other blocks. Batches comprising multiple types are indicated by multiple colors. B: Fraction of pairs starting in a Promoter, Exon, or Other (non-promoter, non-exon) region.

One of the most striking differences between the HeliScopeCAGE-based FANTOM5 CAGE data and the nanoCAGE-based FANTOM5 CAGEscan data is a larger amount of start sites in the gene body, far from the promoter. This can be explained by the lower stringency of the nanoCAGE protocol, which uses template-switching for capturing 5′ ends from limiting amounts of samples [6], where the HeliScopeCAGE protocol, that uses CAP Trapper [19], would not be possible. Readers curious about the position of the random priming site, indicated by the end position of the CAGEscan pairs, will notice that their distribution is very far from random. Control experiments performed using different batches of random primers ordered by different makers confirmed that the quality of the oligonucleotides was not in question (data not shown). In the latest version of the nanoCAGE protocol [20], this problem was solved by the fragmentation of the cDNAs by the "tagmentation" method. Altogether, we recommend to use our latest protocol for making new libraries.

In this study, the CAGEscan libraries were prepared using the nanoCAGE method, but the CAGEscan workflow, which can use any paired-end sequencing of CAGE libraries were the 3′ sequencing read is at a random position in the cDNA, can be applied to other publicly available dataset, for instance made with the RAMPAGE method [21].

Acknowledgments

FANTOM5 was made possible by research grants for the RIKEN Omics Science Center and the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT to Y.H. It was also supported by research grants for the RIKEN Preventive Medicine and Diagnosis Innovation Program (RIKEN PMI) to Y.H. and the RIKEN Centre for Life Science Technologies, Division of Genomic Technologies (RIKEN CLST (DGT)) from the MEXT, Japan. A.R.R.F. is supported by a Senior Cancer Research Fellowship from the Cancer Research Trust, the MACA Ride to Conquer Cancer and the Australian Research Council's Discovery Projects funding scheme (DP160101960). We thank RIKEN GeNAS for generation of the CAGEscan libraries, the Netherlands Brain Bank for brain materials, and the RIKEN BioResource Centre for providing cell lines.

Competing financial interests

The author(s) declare no competing financial interests.

References

[1] Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences* **100**, 15776–15781 (2003). URL `http://www.pnas.org/cgi/doi/10.1073/pnas.2136655100`.

[2] Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics* **38**, 626–635 (2006).

[3] Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014). URL `http://www.nature.com/doifinder/10.1038/nature13182`.

[4] Arner, E. *et al.* Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347**, 1010–1014 (2015). URL `http://www.sciencemag.org/cgi/doi/10.1126/science.1259418`.

[5] Noguchi, S. & consortium, T. F. FANTOM5 CAGE profiles of human and mouse samples (2017).

[6] Plessy, C. *et al.* Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nature Methods* **7**, 528–534 (2010).

[7] Kratz, A. *et al.* Digital expression profiling of the compartmentalized translatome of Purkinje neurons. *Genome Research* **24**, 1396–1410 (2014).

[8] Severin, J. *et al.* Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. *Nature Biotechnology* **32**, 217–219 (2014).

[9] Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology* **16**, 22 (2015).

[10] Abugessaisa, I. *et al.* FANTOM5 transcriptome catalog of cellular states based on Semantic MediaWiki. *Database: The Journal of Biological Databases and Curation* **2016** (2016).

[11] Salimullah, M., Sakai, M., Mizuho, S., Plessy, C. & Carninci, P. NanoCAGE: a high-resolution technique to discover and interrogate cell transcriptomes. *Cold Spring Harbor Protocols* **2011**, pdb.prot5559 (2011).

[12] Tang, D. T. P. *et al.* Suppression of artifacts and barcode bias in high-throughput transcriptome analyses utilizing template switching. *Nucleic Acids Research* **41**, e44–e44 (2013). URL `http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gks1128`.

[13] Hasegawa, A., Daub, C., Carninci, P., Hayashizaki, Y. & Lassmann, T. MOIRAI: a compact workflow system for CAGE analysis. *BMC bioinformatics* **15**, 144 (2014).

[14] Mizuno, Y. *et al.* Increased specificity of reverse transcription priming by trehalose and oligo-blockers allows high-efficiency window separation of mRNA display. *Nucleic Acids Research* **27**, 1345–1349 (1999).

[15] Lassmann, T., Hayashizaki, Y. & Daub, C. O. TagDust–a program to eliminate artifacts from next generation sequencing data. *Bioinformatics (Oxford, England)* **25**, 2839–2840 (2009).

[16] Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754–1760 (2009).

[17] Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078–2079 (2009).

[18] Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).

[19] Kanamori-Katayama, M. *et al.* Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Research* **21**, 1150–1159 (2011).

[20] Poulain, S. *et al.* NanoCAGE: A Method for the Analysis of Coding and Noncoding 5'-Capped Transcriptomes. *Methods in Molecular Biology (Clifton, N.J.)* **1543**, 57–109 (2017).

[21] Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Research* **23**, 169–180 (2013).

Data Citations

1. FANTOM5 CAGEscan view on the ZENBU genome browser: http://fantom.gsc.riken.jp/zenbu/gLyphs/#config=ZkJi4RdBAFhnsudxePrZxD

2. DDBJ Sequence Read Archive, DRA005606 (2017).