Enhort: a platform for deep analysis of genomic positions

Michael Menzel [1], **Peter Koch** [1], **Stefan Glasenhardt** [1], **Andreas Gogol-Döring** [Corresp. 1]

[1] MNI, Technische Hochschule Mittelhessen - University of Applied Sciences, Giessen, Hessen, Germany

Corresponding Author: Andreas Gogol-Döring
Email address: andreas.gogol-doering@mni.thm.de

The rise of high-throughput methods in genomic research greatly expanded our knowledge about the functionality of the genome. At the same time, the amount of available genomic position data increased massively, e.g., through genome-wide profiling of protein binding, virus integration or DNA methylation. However, there is no specialized software to investigate integration site profiles of virus integration or transcription factor binding sites by correlating the sites with the diversity of available genomic annotations.

Here we present Enhort, a user-friendly software tool for relating large sets of genomic positions to a variety of annotations. It functions as a statistics based genome browser, not focused on a single locus but analyzing many genomic positions simultaneously. Enhort provides comprehensive yet easy-to-use methods for statistical analysis, visualization, and the adjustment of background models according to experimental conditions and scientific questions.

Enhort is publicly available online at enhort.mni.thm.de and published under GNU General Public License.

Enhort: A Platform for Deep Analysis of Genomic Positions

Michael Menzel[1]**, Peter Koch**[1]**, Stefan Glasenhardt**[1]**, and Andreas Gogol-Döring**[1]

[1]**MNI, Technische Hochschule Mittelhessen - University of Applied Sciences, Giessen, Hessen, Germany.**

Corresponding author:
Andreas Gogol-Döring[1]

Email address: andreas.gogol-doering@mni.thm.de

ABSTRACT

The rise of high-throughput methods in genomic research greatly expanded our knowledge about the functionality of the genome. At the same time, the amount of available genomic position data increased massively, e.g., through genome-wide profiling of protein binding, virus integration and DNA methylation. However, there is no specialized software to investigate integration site profiles of virus integration or transcription factor binding sites by correlating the sites with the diversity of available genomic annotations. Here we present Enhort, a user-friendly software tool for relating large sets of genomic positions to a variety of annotations. It functions as a statistics based genome browser, not focused on a single locus but analyzing many genomic positions simultaneously. Enhort provides comprehensive yet easy-to-use methods for statistical analysis, visualization, and the adjustment of background models according to experimental conditions and scientific questions.
Enhort is publicly available online at enhort.mni.thm.de and published under GNU General Public License.

INTRODUCTION

Some viruses like HIV (Craigie and Bushman, 2012) and AAV (Deyle and Russell, 2009) are able to copy their genomic sequence into the genome of an infected cell. This can have severe impact on host cell stability as the integration may hit and disable a gene or a regulatory region. The investigation of characteristics and underlying driving factors for virus integration is not only relevant for virology and infectious diseases research but also for approaches in gene therapy that apply virus-derived vectors and transposons to deliver functional DNA fragments into host cells (Riviere et al., 2012; Li et al., 2015). Each gene delivery system has its own mechanisms for genomic integration and preferences for choosing integration sites, hence different systems may have different risks for causing undesired side effects.

Next Generation Sequencing (NGS) facilitates the genome-wide profiling of integration sites, as they are collected e.g. in investigations of protein binding, virus/transposon integration or DNA methylation. Integration sites are available from databases like the Retrovirus Integration Database (Shao et al., 2016) and are regularly created for novel targeted vectors. Typically, the identified sites are related to a variety of genomic features and any integration preferences are determined by a comparison of actual integration sites to a set of random control sites (Gogol-Döring et al., 2016). A proper background model should mimic all known biases of the signal data originating from experimental or laboratory conditions. If, for example, a profiling method is only capable of detecting integration events that are close to certain enzyme restriction sites then the control sites should also be selected accordingly.

Several tools have been published that are capable of processing genomic positions and annotations, like the Genomic HyperBrowser (Sandve et al., 2013). Genome browsers like the UCSC Genome Browser (Kent et al., 2002), IGV (Robinson et al., 2011) or Artemis (Carver et al., 2011) are designed for inspecting single genomic locations. Also custom written scripts are commonly used for the analysis of genomic positions (Cook et al., 2014) or libraries like PyBedTools (Janovitz et al., 2014; Dale et al., 2011). Once written these scripts have the benefit of being a reusable option to conduct a specific set of analysis on recurring data. However, they are limited by the available functionality because each function has be
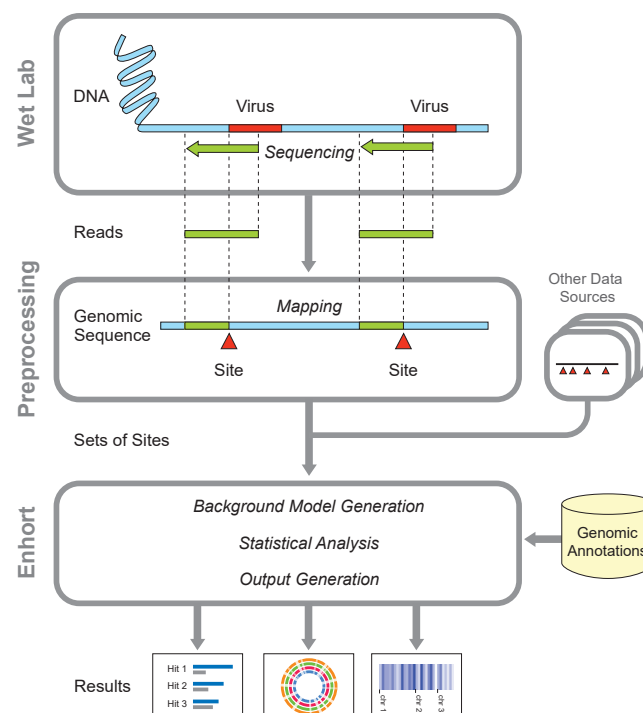
Figure 1. Overview of preparatory work and data gathering for analysis in Enhort. Reads containing viral integration sites are identified and sequenced in the WebLab and mapped to a reference genome. Identified insertion sites are converted to a BED file for the usage in Enhort. Together with genomic annotations from public database the analysis in Enhort is conducted to generated analysis of the given integration sites.

newly developed. Additionally, comparability across laboratories is afflicted by varying functionality and different implementations of background models. There is yet no specialized tool for genomic positions analysis that combines the features of instant analysis and user defined adaptable background models that mimic known biases.

In this paper we present Enhort, a user-friendly web-platform for deep analysis of large sets of genomic positions. Our aim is to accelerate and simplify the data analysis process as well as to standardize it in order to increase reproducibility. Enhort is capable of adjusting background sites used for comparison by user selected covariates. This includes annotation tracks like restriction sites or chromatin accessibility, gene expression tracks and sequence motifs. With covariates it is possible to adjust the background sites selection in a way that they match the investigated sites for a specific track. The adaptation rules out the effects of this annotation for the background. This feature can be used to adjust for experimental bias as well as specific questions. Figure 1 shows the schematic process of data gathering and the usage of Enhort in the workflow of analyzing genomic positions.

METHODS

Integration sites of viruses are gathered by sequencing infected cells and preprocessing as shown in Figure 1. These sites are uploaded to Enhort and are intersected with each annotation file to compute fold-change enrichment and χ^2 test in comparison to control sites, yielding a measure for effect strength and significance of each annotation respectively. Figure 2 shows the schematic analysis pathway for sites uploaded by a user. Statistical analysis depends on the Apache Commons Math library [1] and uses Bonferroni correction for multiple hypothesis testing. The libraries plotly.js [2] and Circos [3] are used for visualization. The results are sorted according to their relevance and presented in conjunction with

---

[1] https://commons.apache.org/proper/commons-math/
[2] https://plot.ly/javascript/
[3] http://circos.ca/
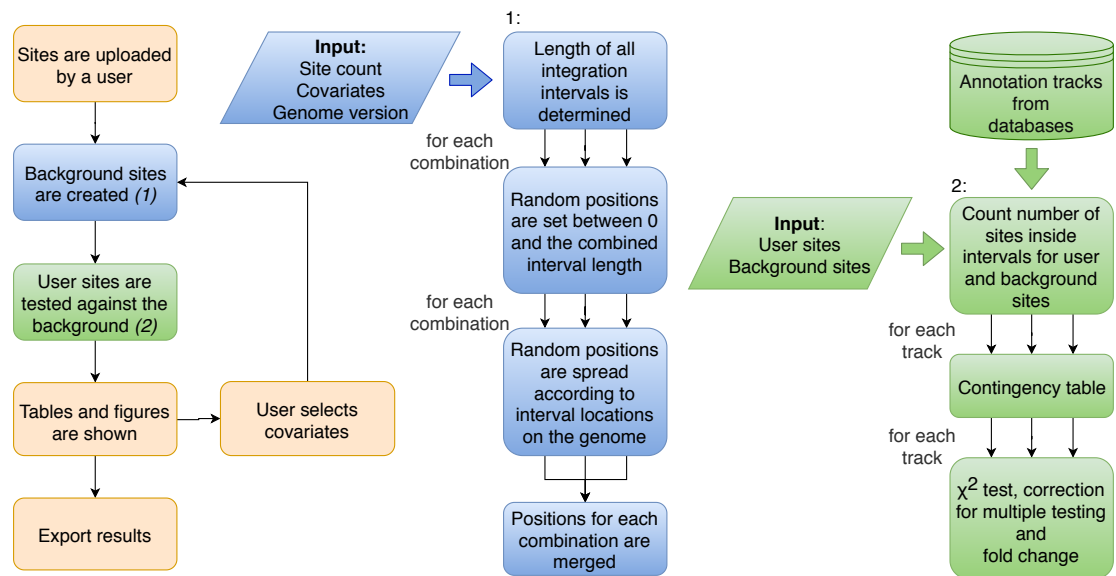
Figure 2. Flowchart of the procedure of analysis performed by Enhort. Blue boxes show the steps to create a background model based on multiple covariates. Random positions have to be set for each combination of covariates. Green boxes show the steps to test the user sites against the background sites. The results are returned as a table and converted into figures for the user.

appropriate figures. Example results for a virus can be seen in Figure 3A. The software has been designed in a way that analysis results are almost immediately available after upload.

In many cases a background model consisting of random sites is not sufficient for an adequate analysis. Some protocols, for example, can only detect integration events that occurred in close proximity to a restriction site of a specific enzyme, like EcoRI, which cuts inside of GAATTC hexamers (Pingoud and Jeltsch, 2001). Background models should be adapted to mimic the actual integration pattern with regard to any known technical bias. In this case, the control sites should also be selected to be near restriction sites. This can be achieved in Enhort by setting the appropriate genome annotation as a covariate. When selecting the track that contains all possible genomic positions of GAATTC hexamers as covariate, Enhort will generate a set of control sites having exactly the same distribution of distances to the enzyme restriction sites as the actual virus integration sites.

Covariates help to adapt the background model both for technical circumstances, for example, restriction sites and for eliminating a bias or biological preferences such as motifs or genetic features. Covariates can also be used to identify dependent or separate weak integration preferences that are covered by stronger effects, as shown in Figure 3 B. MLV integration sites are compared to two different control sets: A random and an altered background, to identify the actual integration preferences; e.g. for histone mark H3K4me3, which is a known preference of MLV (Gogol-Döring et al., 2016).

For the validity of statistical testing it is usually indispensable to normalize the background model relative to multiple covariates. For that purpose, Enhort supports the selection of multiple covariates simultaneously in order to further investigate the integration site characteristics. For example, Enhort may create a control set that considers chromatin accessibility, restriction site distance as well as several histone modifications simultaneously. This functionality is needed to build background models for sites that are influenced by multiple factors, e.g. biological and technical biases. A set of additional features listed in the following table:

1. Statistical analysis for annotation tracks:
   (a) Fold change
   (b) χ^2 test
   (c) Kolmogorov-Smirnov test
2. Hotspot analysis (Figure 4 C)
3. Position depended enrichment (Figure 4 A)
4. Background models based on:
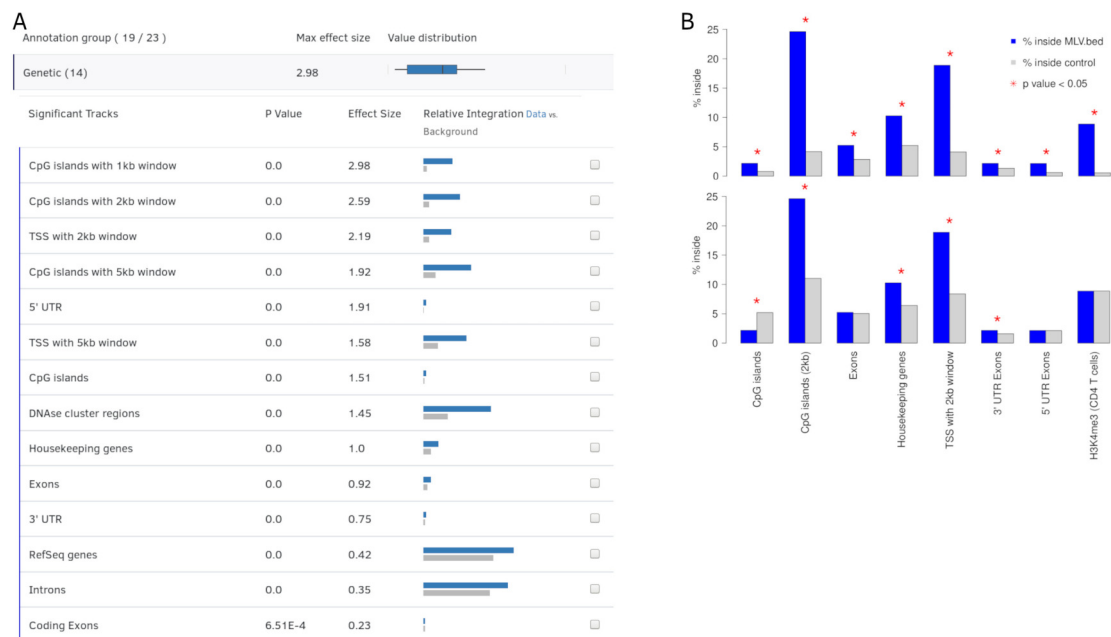
Figure 3. **A:** Output view example, generated by Enhort when analyzing Murine Leukemia Virus (MLV) integration sites in CD4$^+$ T cells (Roth et al., 2011). The results are presented in a table containing for each annotation the p value, effect size and a visual representation of the integration. The annotations are ranked by effect strength. **B:** Effect of covariate selection. The upper diagram contains integration frequencies of MLV compared to random sites for a selection of annotations. This virus is known for preferentially integrating near transcription start sites (TSS) and H3K4me3 histone marks (LaFave et al., 2014). The lower diagram shows the same data after selecting H3K4me3 as covariate. The adapted background model is generated in a way that control sites and MLV integration sites have the same frequency relative to H3K4me3. This also changed the control site frequencies for other annotations: MLV integration is no longer enriched but depleted in CpG islands when compared to the adapted background model.

99   (a) Inside and outside of annotations
100   (b) Distance to annotations
101   (c) Scored annotations
102   (d) Sequence logo
103   5. Upload background sites
104   6. Comparing effects of different background models
105   7. Batch analysis of multiple integration sets
106   8. Heatmaps to compare integration sets (Figure 4 B)
107   9. Custom annotation tracks
108   10. Blend annotation tracks
109   11. Export results as R code and CSV files

110   Enhort is separated into a lightweight, web-based user interface and a high performance back-end
111 server attached to a SQLite database storing meta-information about the annotations fetched from
112 DeepBlue (Albrecht et al., 2016). Results from Enhort are instantaneously available as seen in table 1
113 where the run times for different input sizes are shown. Our application currently offers 1402 annotation
114 tracks from 97 cell lines and tissues for human genome assemblies hg19 and hg38, downloaded from
115 UCSC Genome Browser (Fujita et al., 2011), Encode (ENCODE Project Consortium, 2004), ChIP-
116 Atlas (*http://chip-atlas.org*), BLUEPRINT Epigenome (Adams et al., 2012) and Roadmap Epigenomics
117 (Roadmap Epigenomics Consortium et al., 2015) using the DeepBlue Epigenomic Data Server (Albrecht
118 et al., 2016).
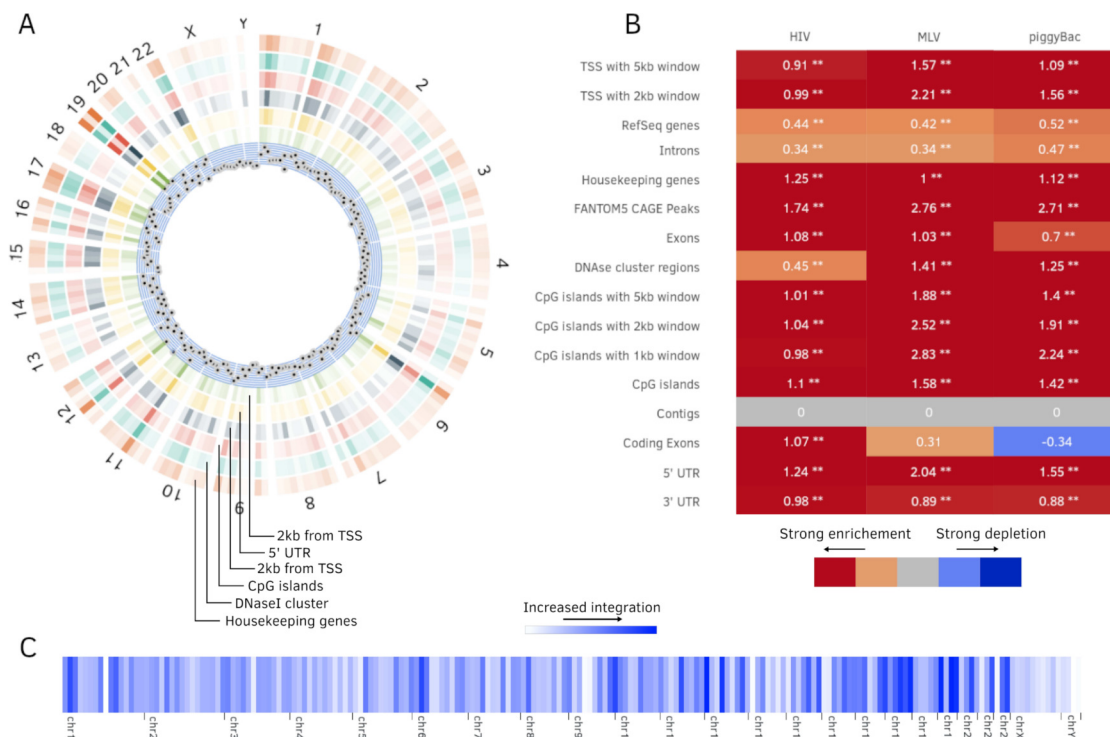
Figure 4. Additional plots generated by Enhort. **A:** Circos plot (Krzywinski et al., 2009) of position dependent enrichment over all chromosomes for MLV for the most significant tracks. **B:** Heatmap for a set of three integration data sets against various annotations. The values are $\log_2$-fold changes of the numbers of integration vs control sites falling into a given annotation. Star symbols mark statistically significant changes. The same background sites are used for the comparisons. The background sites are adapted to integrate only inside the sequence contigs. **C:** Integration hotspots across the genome for MLV. The color intensity of the thin bars show the integration ratio inside of the respective genomic region.

RESULTS AND DISCUSSION

Literature Review

We reviewed the relevance of Enhort for contemporary research by systematically searching PubMed, Google Scholar, and several review articles for publications concerning the analysis of genomic integration sites. The publications include virus integration site analysis for HIV, MLV, HRP-2, SIV, foamy virus, HPV, AAV and transposons such as piggyBac, LINE-1, Alu and sleeping beauty. In total we identified 59 relevant publications. Details on the reviewed publications and methodological analysis are available in the supplementary table S1. Of these publications 19 used completely random control sites, only 6 used adapted control sites. The data analyses presented in 37 (63%) publications could have been entirely performed with our tool. 6 further publications use at least some methods provided by Enhort. We assume that if they had the opportunity to use Enhort the authors would have saved a lot of effort writing custom analysis scripts.

Data Re-analysis

To further present the capabilities of Enhort we re-analyzed integration sites of the PiggyBac transposon (PB) published by Gogol-Döring et al. (2016) using Enhort. Results from Wilson et al. (2007) are used for comparison. PB integration characteristics show a preference for genes, exons, introns, highly expressed genes, DNase I hypersensitive sites, H3K4me3 and open chromatin structures (Wilson et al., 2007) (Li et al., 2013). We uploaded the PB integration sites to Enhort, selected all relevant tracks and finally exported the results. Figure 5 A shows the log fold changes for a selection of annotations for PB against a random background in grey. Figure 5 B shows the sequence logos for the PB integration sites and the random background. The barplots were created using the R-export feature of Enhort.

Track count	23			1127		
Covariate count	0	2	5	0	2	5
Site count	Execution time (ms)					
150k	877	1188	4668	8538	10436	12540
125k	717	1103	5628	5509	7552	7975
100k	749	817	5085	3724	4672	8673
75k	624	571	4019	4905	4397	9633
50k	470	555	5455	4736	5844	10451
25k	308	351	4628	3246	3091	8111

Table 1. Analysis execution times for different usual site counts, annotation tracks from hg19 and covariate counts. (Back-end server: SuperMicro SuperServer 4048B-TRFT 4x Intel Xeon E7-8867v3 with 2048GB DDR3 ECC LR)
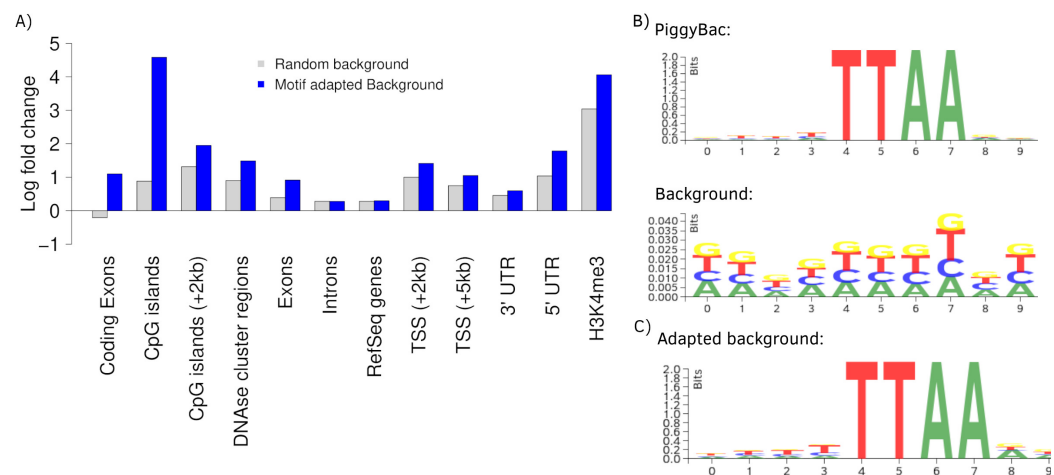
Figure 5. A Log fold changes of PB integration sites in relation to several annotations against a random and an adapted background model. Changing the background model to adapt the TTAA motif changes the observation of several integration preferences. **B** The PB motif and random sites motif, corresponding with the random background bars in **A**. **C** Motif of the random sites after adaption to the PB motif using Enhort.

The key feature of the PB integration preference is the TTAA motif in which all integrations occur. To precisely analyze the preferences of PB integration the background model has to be adapted to replicate the TTAA motif preference. This can be achieved using Enhort by creating a set of pseudo-random control sites that are located only inside a TTAA sequence. To achieve this, we simply selected the sequence logo as a covariates. Enhort takes genomic positions from a pre-sampled set of positions where each position has a probability based on the similarity between the surrounding sequence and the TTAA sequence. The results are shown in Figure 5 C where the background sites and PB show a similar motif after the motif is added as a covariate using Enhort. The motif adaption also changes the observed integration characteristics seen in Figure 5 A. The relative decreased integration of PB into coding exons is changed to a significant preference, because CpG islands are less likely to be hit by a site from the adapted background model, as TTAA occurs relatively less frequent in CpG islands. The same applies to DNAse cluster regions, TSS and exons, where the significance of integration is enriched in comparison to a random background. Only a small change for the enrichment in introns and genes is visible. Overall this indicates that beside the TTAA preference of PB there are additional mechanisms that alter the integration preferences. Using the background adaption feature of Enhort it would be possible to test different hypothesis against the data and build a model that explains the integration preferences.

To further review the analytic capabilities of our software, the integration counts of PB sites are compared to published results from Wilson et al. (2007). The comparison can be seen in Table 2. An

158 increased integration of PB into RefSeq genes, inside the 5kb-TSS window, as well as a preference for
159 CpG islands is observable for both analyses.

Annotation track	Enhort Fold change	Wilson et al. Fold change	Enhort PB (%)	Wilson et al. PB (%)	Enhort Random (%)	Wilson et al. Random (%)
RefSeq genes	1.32	1.46	63.08	48.8	47.93	33.2
TSS (±5 kb)	2.14	3.00	20.8	16.2	9.7	5.4
CpG islands (±1 kb)	5.52	2.00	12.99	3.8	2.35	1.9
CpG islands (±5 kb)	2.82	0.96	22.85	7.7	8.09	8.3
Repeats:						
LINE	0.71	0.76	7.72	12.7	10.90	16.7
SINE	0.50	0.54	3.8	6.0	7.64	11.1
LTR	0.56	1.84	2.79	6.8	5.0	3.7
DNA	1.61	1.18	1.87	4.0	1.61	3.4

Table 2. Log fold changes and integration ratios of Wilson et al. (2007) in comparison to Enhort for two PB integration site sets.

160 Wu et al. (2003) published a study on MLV and HIV stating that MLV favors TSS regions, whereas
161 HIV does not display a strong preference towards TSS regions. The available integration sites were
162 uploaded to Enhort and analyzed using the batch tool with a random 10,000 site background model. The
163 results from Enhort show a similar integration pattern as stated in Wu et al. (Table 3). Except for CpG
164 islands for HIV where Wu et al. found a near random integration and we found a decreased integration.
165 For further review, HIV and MLV integration sites were uploaded independently to Enhort, and
166 RefSeq genes added as covariate. This background model had only a little effect on MLV as the preference
167 for TSS and CpG islands only changed slightly, indicating that the preference for TSS is not due to a
168 preference for RefSeq genes. For the HIV integration sites the housekeeping genes, which are a known
169 preference of HIV (Craigie and Bushman, 2012), are still statistically significant against this background
170 model.
171 Finally, RefSeq genes and TSS (± 5kb) were both used as covariates together, showing that the
172 integration ratio of MLV into CpG islands with a (± 1kb) window decreases slightly. This shows that
173 the integration into the CpG islands is probably not a side effect of the preference for TSS or genes. The
174 combined background model with RefSeq genes and TSS does not have any influence on the HIV fold
175 changes compared to the previous background model.
176 The creation of each background model and comparing the results was possible using built-in features
177 of Enhort. We further added histone modifications to the analysis showing that H4K20me1 is significantly
178 enriched for both integration sets and does not change significantly for the different background models.
179 This indicates that the histone modification preferences is an additional effect, only slightly influenced by
180 the preference for genes and TSS. H3K4me2 and H3K27ac are known preferences of MLV (De Ravin
181 et al., 2014) and show a high fold change for all background models. With the available database it would
182 be easy to add numerous additional annotations for comparison.
183 We have shown that Enhort is capable of reproducing integration site analysis with less effort
184 and additionally offers easy-to-use mechanisms to create more sophisticated analysis using adaptable
185 background models. The exact annotation files were not available for comparison, so it was not possible
186 to produce the exact numbers. However, Enhort uses the same calculation principle. With the same
187 annotations and sites the results by Enhort would be the same as in the referenced publications.

CONCLUSION

189 In this publication we present Enhort, a fast and easy-to-use analyzing platform for genomic positions.
190 Based on a comprehensive library of genomic annotations, Enhort provides a wide range of methods to
191 analyze large sets of sites. In contrast to multi-purpose software such as bioconductor, Enhort enables
192 scientists to analyze data without programming effort or extensive manual work.
193 Our literature review shows that Enhort is able to perform most of the analyses commonly used in
194 the investigation of integration sites. The re-analysis of Wilson et al. and Wu et al. demonstrates that

	Wu et al.		Enhort					
	HIV	MLV	HIV	MLV	HIV[†]	MLV[†]	HIV[‡]	MLV[‡]
RefSeq genes	2.58*	1.5*	1.7*	1.4*	1	1	1	1
Housekeeping genes	-	-	3.7 *	1.36	2.22 *	1.12	2.05*	1.04
CpG islands (± 1 kb)	1	8*	0.41	6.24*	0.35	6.17*	0.31	4.09*
TSS (± 5 kb)	2.5*	4.7*	1.34	2.3*	1.14	2.02*	1	1
H4K20me1	-	-	1.71*	1.56*	1.34*	1.52*	1.36*	1.42*
H3K4me2	-	-	1.23	21.7*	1.48	21.29*	1.09	15.2*
H3K27ac	-	-	0.9	24.52*	1.01	22.79*	0.83	20.12*

Table 3. Comparison between fold changes of Wu et al. (2003) and Enhort over different annotations on the same integration sites. * $P < 0.002$, [†] with RefSeq genes as covariate, [‡] with RefSeq genes and TSS ($\pm$ 5kb) as covariates.

Enhort is able to reproduce analyses from literature with little effort. It was not possible to reproduce the exact values, because the version of the annotation was not recorded in the publications. However, more detailed insights can be made using adaptable background models. This was shown in the comparison of HIV and MLV from Wu et al. against different control sites.

Most publications use very simple background models for statistical analysis of integration data and could potentially be improved using better background models. Enhort provides methods to easily create more sophisticated background models for improving both the accuracy and the range of possible analyses. Complex background models can be used to identify weak effects and segregate driving factors for integration, find a minimal set of annotations to mimic integration characteristics, as well as to eliminate technical biases. In conclusion, this shows that Enhort will be a valuable tool for further analyses of genomic positions, no matter if these positions are derived from virus integration, sequence motifs, enzyme restrictions, histone modifications, or protein binding.

REFERENCES

Adams, D., Altucci, L., Antonarakis, S., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., Dahl, F., Dermitzakis, E., Enver, T., Esteller, M., Estivill, X., Ferguson-Smith, A., Fitzgibbon, J., Flicek, P., Schacht, C., and Willcocks, S. (2012). Blueprint to decode the epigenetic signature written in blood. *Nature biotechnology*, 30(3):224–226.

Albrecht, F., List, M., Bock, C., and Lengauer, T. (2016). Deepblue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. *Nucleic acids research*, 44(W1):W581–W586.

Carver, T., Harris, S. R., Berriman, M., Parkhill, J., and McQuillan, J. A. (2011). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, 28(4):464–469.

Cook, L. B., Melamed, A., Niederer, H., Valganon, M., Laydon, D., Foroni, L., Taylor, G. P., Matsuoka, M., and Bangham, C. R. (2014). The role of htlv-1 clonality, proviral structure, and genomic integration site in adult t-cell leukemia/lymphoma. *Blood*, 123(25):3925–3931.

Craigie, R. and Bushman, F. D. (2012). Hiv dna integration. *Cold Spring Harbor perspectives in medicine*, 2(7):a006890.

Dale, R. K., Pedersen, B. S., and Quinlan, A. R. (2011). Pybedtools: a flexible python library for manipulating genomic datasets and annotations. *Bioinformatics*, 27(24):3423–3424.

De Ravin, S. S., Su, L., Theobald, N., Choi, U., Macpherson, J. L., Poidinger, M., Symonds, G., Pond, S. M., Ferris, A. L., Hughes, S. H., HL, M., and X., W. (2014). Enhancers are major targets for murine leukemia virus vector integration. *Journal of virology*, 88(8):4504–4513.

Deyle, D. R. and Russell, D. W. (2009). Adeno-associated virus vector integration. *Current opinion in molecular therapeutics*, 11(4):442.

ENCODE Project Consortium (2004). The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640.

Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T. R., Giardine, B. M., Harte, R. A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R. M., Learned, K., Li, C. H., Meyer, L. R., Pohl, A., Raney, B. J., Rosenbloom, K. R., Smith, K. E., Haussler, D., and Kent, W. J. (2011). The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*, 39.

Gogol-Döring, A., Ammar, I., Gupta, S., Bunse, M., Miskey, C., Chen, W., Uckert, W., Schulz, T. F., Izsvák, Z., and Ivics, Z. (2016). Genome-wide profiling reveals remarkable parallels between insertion site selection properties of the mlv retrovirus and the piggybac transposon in primary human cd4+ t cells. *Molecular Therapy*, 24(3):592–606.

Gogol-Döring, A., Ammar, I., Gupta, S., Bunse, M., Miskey, C., Chen, W., Uckert, W., Schulz, T. F., Izsvák, Z., and Ivics, Z. (2016). Genome-wide Profiling Reveals Remarkable Parallels Between Insertion Site Selection Properties of the MLV Retrovirus and the piggyBac Transposon in Primary Human CD4+ T Cells. *Molecular Therapy*, 24(3):592–606.

Janovitz, T., Oliveira, T., Sadelain, M., and Falck-Pedersen, E. (2014). Highly divergent integration profile of adeno-associated virus serotype 5 revealed by high-throughput sequencing. *Journal of virology*, 88(5):2481–2488.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at ucsc. *Genome research*, 12(6):996–1006.

Krzywinski, M. I., Schein, J. E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome research*.

LaFave, M. C., Varshney, G. K., Gildea, D. E., Wolfsberg, T. G., Baxevanis, A. D., and Burgess, S. M. (2014). MLV integration site selection is driven by strong enhancers and active promoters. *Nucleic Acids Research*, 42(7):4257–4269.

Li, L., Zhang, D., Li, P., Damaser, M., and Zhang, Y. (2015). Virus integration and genome influence in approaches to stem cell based therapy for andro-urology. *Advanced Drug Delivery Reviews*, 82-83:12–21.

Li, M. A., Pettitt, S. J., Eckert, S., Ning, Z., Rice, S., Cadiñanos, J., Yusa, K., Conte, N., and Bradley, A. (2013). The piggyBac Transposon Displays Local and Distant Reintegration Preferences and Can Cause Mutations at Noncanonical Integration Sites. *Molecular and Cellular Biology*, 33(7):1317–1330.

Pingoud, A. and Jeltsch, A. (2001). Structure and function of type ii restriction endonucleases. *Nucleic acids research*, 29(18):3705–3727.

Riviere, I., Dunbar, C. E., and Sadelain, M. (2012). Hematopoietic stem cell engineering at a crossroads. *Blood*, 119(5):1107–1116.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. *Nature biotechnology*, 29(1):24.

Roth, S. L., Malani, N., and Bushman, F. D. (2011). Gammaretroviral Integration into Nucleosomal Target DNA In Vivo. *Journal of Virology*, 85(14):7393–7401.

Sandve, G. K., Gundersen, S., Johansen, M., Glad, I. K., Gunathasan, K., Holden, L., Holden, M., Liestøl, K., Nygård, S., Nygaard, V., Paulsen, J., Rydbeck, H., Trengereid, K., Clancy, T., Drabløs, F., Ferkingstad, E., Kalaš, M., Lien, T., Rye, M. B., Frigessi, A., and Hovig, E. (2013). The Genomic HyperBrowser: an analysis web server for genome-scale data. *Nucleic Acids Research*, 41(W1):W133–

286 W141.

287 Shao, W., Shan, J., Kearney, M. F., Wu, X., Maldarelli, F., Mellors, J. W., Luke, B., Coffin, J. M., and

288 Hughes, S. H. (2016). Retrovirus integration database (rid): a public database for retroviral insertion

289 sites into host genomes. *Retrovirology*, 13(1):47.

290 Wilson, M. H., Coates, C. J., and George, A. L. (2007). PiggyBac Transposon-mediated Gene Transfer in

291 Human Cells. *Molecular Therapy*, 15(1):139–145.

292 Wu, X., Li, Y., Crise, B., and Burgess, S. M. (2003). Transcription start regions in the human genome are

293 favored targets for mlv integration. *Science*, 300(5626):1749–1751.