A Simulation Study of Patient Accrual Patterns in Clinical Trials and Data Analysis of Histone 3 Lysine 36 Trimethylation ChIP-seq in Human Kidney Cancer

A Thesis
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Shichao Yu

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Dr. Karla Ballman, advisor
Dr. Huihuang Yan, co-advisor

July 2017

Acknowledgements

I gratefully acknowledge my advisor, Dr. Karla Ballman for her advice and guidance on the simulation study. She provided me a thorough research plan and allowed me to expand her previous research. My sincere gratitude goes to Dr. Huihuang Yan at Mayo Clinic who joined to my committee and served as co-advisor on ChIP-Seq data analysis. He helped me every step on the uses of applications, programming and study design. I would like to thank Dr. Claudia Neuhauser, advisor on Master program who provided me the access to the high performance computational platform at UMN and made it possible to continue my research.

It is pleasure to thank those who made my thesis possible.

Dedication

This thesis is dedicated to my family

Abstract

In part one, we simulated a successive of two-armed randomized clinical trial with the time-to-event outcome over 15 years. We used three different accrual pattern representing slow, medium and fast accrual, which is in fact related to the number of trials for the sequential trials interested in the 15-year period. We used a historical survival distribution to explore the treatment effects and analyzed by the Cox proportional hazard ratio model and log-rank test. We computed the mean and median overall hazard ratio (year 15 versus year 0), and the probability of detrimental effect to find the optimal design parameters. Finally, we carried out a sensitivity analysis to study the effect of an additional 6 month turnaround time.

In Part two, we have described a general workflow for the normalization of ChIP-seq data by estimating the normalization factor from peak-less regions. Using publicly available histone 3 lysine 36 trimethylation ($H_3K_{36}me_3$) data from human kidney cancer, we demonstrated the better performance of our method over the existing approach.

Table of Contents

List of Tables

List of Figures

Part I A Simulation Study of Patient Accrual Patterns in Clinical Trials

I-1 INTRODUCTION

Clinical trials have played an important role in the development of drugs or treatments, which is a time consuming and expensive process.[1] It usually involves recruiting subjects, designing treatment strategy, collecting data and evaluating the results. Sample size estimation and recruitment are the most fundamental parts of the research design and determine the success of the trial to a large extent. It has been reported that clinical trials fail due to inefficient and insufficient patient recruitment process.[2]

Survival analysis compares the treatment effects between the control and the experimental groups at multiple points in time.[3] A special case is the oncology clinical trial. For this kind of disease, it is more difficult to recruit enough people than for other diseases over a long-term research period.[4] A small gain in survival is considered as clinically relevant. To distinguish this subtle difference, a much larger sample size is required. These two factors make clinical trials with large sample size less practical. In response to the complexity of the clinical trial design and the occurrence of clinical trial failures, clinical trial simulation (CTS) based on known knowledge has been applied to model a clinical trial under the different trial scenarios and assess the probability of a successful outcome since the late twentieth century.[5] The merit of the CTS lies in its high efficiency, low cost and flexible adjusting of different trial scenarios and related design parameters for unlimited number of trials with reliable knowledge inherited from

historical clinical trial analysis.[6] For example, Strauss and Simon proposed a process for clinical trials that has a limited number of patients available for treatments over a period. Instead of a single trial, they considered to carry out a sequential two-armed randomized clinical trials. In the end, they choose an optimal treatment based on the expected success probability.[7] Later Sposto and Stram pioneered a simulation on pediatric cancer trials with smaller sample sizes in a survival analysis setting.[8] They investigated the treatment efficacy of factors like trial duration, significance level, patient accrual rate, and sample size in a series of two-treatment randomized trials. They found that larger α level and a series of smaller trials in a 25-year research course afforded larger average gains in cure rate. Deley and Ballman further expanded this strategy by carrying out more trials (smaller) in different accrual patterns with different evaluation criteria over a 15-year horizon.[9] Their findings implied that both the significance level and the number of trials (related to the sample size) played a role in the treatment effect when trials with smaller sample size were performed.

In this research, we simulated a successive two-armed randomized clinical trial with the time-to-event outcome over 15 years. We used three different accrual pattern representing slow, medium and fast accrual, which is in fact related to the number of trials for the sequential trials interested in the 15-year period. We used a historical survival distribution to explore the treatment effects and deployed the Cox proportional hazard ratio model and log-rank test for analysis. We computed the mean and median overall hazard ratio (year 15 versus year 0), and the probability of detrimental effect to find the optimal design parameters. Finally, we carried out a sensitivity analysis to study

the effect of an additional 6 month turnaround time. The practical utility of this strategy lies in that with the development of molecular genetics and next generation sequence technology, many cancers originally thought to be the same type now are found having multiple rarer sub-types. Our method provides an insight on how to carry out clinical trials in this new era.

I-2 METHODS

Our current study is based on the theory of survival analysis. We assume that trials are performed in a series of two-treatment randomized clinical trials (control treatment as the current standard of care versus new experimental treatment.). Each trial has an accrual period and a fixed follow-up time (FU) before the start of the next trial, which will be compared with the current standard treatment. The one with a better treatment effect will be used in the next randomized trial. This cycle was repeated until the last trial reached the end of the 15th year with overall survival as the primary endpoint. Simulation parameters include different trial sample sizes, accrual rates, and distributions of treatment effect.

I-2. 1 Simulation parameters and statistical formulation

Sample size *n*

For each trial, the sample size *n* was a constant throughout the simulated successive trials. We will consider ten different sample sizes, which increased from 100 patients to 1000 patients in our trials.

Accrual rate *η*

Accrual rate η is the number of patients accrued per year. We will consider three different accrual rates in our study: 100, 200, and 500 patients/year.

Accrual period T_{Acc}

Accrual period $\boldsymbol{T_{Acc}}$ refers to the length of time required to obtain the desired number of patients to the trial. According to the definition, Accrual period $\boldsymbol{T_{Acc}}$ is given by

$$\boldsymbol{T_{Acc}} = n/\eta \qquad \text{Eq. I-1}$$

Where n is sample size and η is accrual rate.

Follow-up time T_{FU}

Follow-up time $\boldsymbol{T_{FU}}$ means the period starting from the end of accrual of scheduled number of patients for current trial to the beginning of the next one. It consists of the time for the current trial to be finished and analyzed. Generally, a follow up time of one year was considered.

Number of trials *K*

The number of trials K conducted over the 15-year period was derived from the sample size, the expected annual accrual rate, and FU, i.e. the ratio of total research period to the length of each trial (**Figure I-1**). It follows that

$$K = 15/(T_{Acc}+T_{FU}) = 15/(n/\eta+T_{FU}) \qquad \text{Eq. I-2}$$

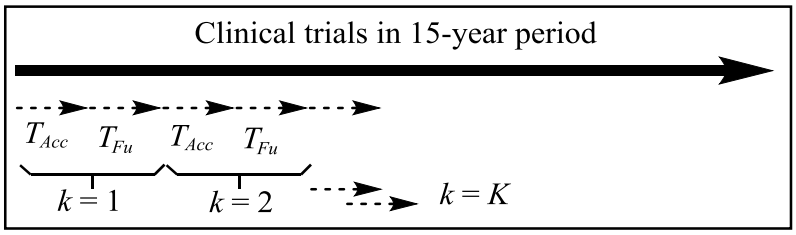

Figure I-1 Number of trials performed in a specific period

Hazard ratio *HR*

The hazard ratio *HR* of the hazard rate for treatment group (λ^E) to the control group (λ^S) is constant over time and derives from a known historical distribution with probability = 10 /[1 +(21.87 × log(HR)^2)] and mean(HR)=0.95.[10]

$$HR_k = \lambda^E_k / \lambda^S_k \qquad k=1, ..., K \qquad \text{Eq. I-3}$$

With a simple transformation from the above equation, we can easily find that the event rate λ^E_k for the experimental treatment in the *k*th trial is proportional to the event rate λ^S_k for the standard treatment in the *k*th trial.

$$\lambda^E_k = HR_k \cdot \lambda^S_k \qquad k=1, ..., K \qquad \text{Eq. I-4}$$

Level of statistical significance, α

α is the probability of wrongly concluding that two treatments A and B differ when in fact they do not differ. In our simulation, we considered the following α values: 2.5%, 5%, 10%, 15%, 20%.

Criteria for treatment selection and parameter evaluation

The experimental treatment will be selected when the following condition is satisfied:

$$HR_k < 1 \quad \text{and} \quad F(\chi^{LR}_{k,I}) < \alpha \qquad \text{Eq. I-5}$$

i.e. the treatment group has better treatment effect than the control group and the probability of the observed log-rank test statistic obtained from trial *k* is less than α. Here HR_k is the estimated hazard ratio for trial *k*. F() denotes the cumulative distribution

function for a *Chi-squared* distribution with 1 degree of freedom (for 2 groups comparison) and χ_k^{LR} is the observed log-rank test statistic obtained from trial *k*.

If the *k*th trial the experimental treatment is selected, it would be the standard treatment for the next trial: $\lambda^s_{k+1} = HR_k \cdot \lambda^s_k$; if not, then the standard treatment in the *k*th trial is retained for the *(k+1)*th trial:

$$\lambda^s_{k+1} = \lambda^s_k \quad \text{Eq. I-6}$$

Let HR^c_{k+1} be the hazard ratio between the standard treatment of the *(k+1)*th trial and the standard treatment of the *k*th trial:

$$HR^c_{k+1} = \lambda^s_k + 1/\lambda^s_k \quad \text{Eq. I-7}$$

Let HR_O be the overall hazard ratio of the treatment event rate after series of *K* trials at the end of the 15 years to the event rate of the standard treatment of the first trial: $HR_O = \lambda^s_{k+1}/\lambda^s_1 = \prod HR^c_{k+1}$. We estimated the expected value of the overall hazard ratio ($E(HR_O)$) by computing the mean of the observed distribution.

The probability of detrimental effect $P(HR_O > 1)$, i.e. the event rate for the treatment selected at the end of the 15 years is worse than the event rate for the initial control group, is applied to evaluate the possibility of harm in the sensitivity analysis.

In our simulation, we will optimize the design parameters by considering the following two criteria: 1) minimization of *E(HRo)*; 2) minimization of *E(HRo)* with the constraint $P[HRo > 1] < 0.025$.

I-2.2 Simulations

I-2.2.1 Scenarios

We considered 2 different disease settings (Scenarios) in 15 years as defined by

expected survival rates and different prevalence rates, which affects the yearly accrual rate (**Figure I-2**).

Scenario 1 (Sc1)

- The baseline survival curve was defined by the λ^s_1 parameter:

Suppose the median survival is 1 year for the standard treatment and the survival distribution for each treatment follows an exponent function, thus we will have:

$$S^s_1(1 \text{ year}) = 0.50 \Rightarrow \quad \lambda^s_1 = -\ln(0.5)/1 \qquad \text{Eq. I-8}$$

- Three accrual rates, η, equal to 100, 200 or 500 patients / year, respectively.

- Sample size increases from 100 to 1000 patients by 100 patients.

- The minimal follow-up duration (T_{FU}) has been set as 1 years.

- Number of trials (K) performed in the 15 years was derived from sample size and accrual rate of each trial (n) according to the following relationship:

$$K = 15/(T_{Acc}+T_{FU}) = 15/(n/\eta+T_{FU}) \qquad \text{Eq. I-9}$$

Scenario 2 (Sc2)

This scenario with 6 months set-up period between two trials is used as sensitivity analysis to compare the treatment effects with scenario 1.

- The baseline survival curve was defined by the λ^s_t parameter:

Suppose the median survival is 1 year for the standard treatment (see Eq. I-10)

- Accrual rates, η, equal to 100 patients / year.

- Sample size equal to 100 patients.

- The follow-up duration (T_{FU}) has been set as 1 year.

- The Set-up time (TSU) i.e. the length between end of a trial and start of the next one,

cannot be ignored and equals to 0.5 year.

- Number of trials (K) performed in the 15 years was derived from sample size and accrual rate of each trial (n) according to the relationship

$$K = 15/(T_{Acc} + T_{SU} + T_{FU}) = 15/(n/\eta + T_{SU} + T_{FU}) \qquad \text{Eq. I-11}$$

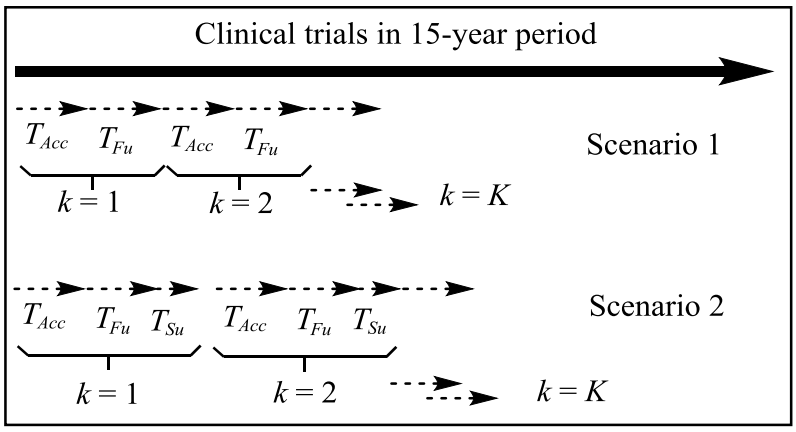

Figure I-2 Scenario 1 and 2 in our simulation

A total of 31 combinations of scenarios, accrual rate, number of trials that could reasonably be run over the 15 years were evaluated (**Table I-1**).

Table I-1 List of the combinations between the baseline situation and the number of trials.

Scenario	Co	Accrual rate η	T_{FU}	S^s_1 (1 year)	Number of patients n	Accrual duration T_{ACC}	Number of trials K
1	1	100	1	0.5	1000	10	1
1	2	100	1	0.5	900	9	1
1	3	100	1	0.5	800	8	1
1	4	100	1	0.5	700	7	1
1	5	100	1	0.5	600	6	2
1	6	100	1	0.5	500	5	2
1	7	100	1	0.5	400	4	3
1	8	100	1	0.5	300	3	3
1	9	100	1	0.5	200	2	5
1	10	100	1	0.5	100	1	7
1	11	200	1	0.5	1000	5	2
1	12	200	1	0.5	900	4.5	2

1	13	200	1	0.5	800	4	3
1	14	200	1	0.5	700	3.5	3
1	15	200	1	0.5	600	3	3
1	16	200	1	0.5	500	2.5	4
1	17	200	1	0.5	400	2	5
1	18	200	1	0.5	300	1.5	6
1	19	200	1	0.5	200	1	7
1	20	200	1	0.5	100	0.5	10
1	21	500	1	0.5	1000	2	5
1	22	500	1	0.5	900	1.8	5
1	23	500	1	0.5	800	1.6	5
1	24	500	1	0.5	700	1.4	6
1	25	500	1	0.5	600	1.2	6
1	26	500	1	0.5	500	1	7
1	27	500	1	0.5	400	0.8	8
1	28	500	1	0.5	300	0.6	9
1	29	500	1	0.5	200	0.4	10
1	30	500	1	0.5	100	0.2	12
2	31	100	1.5	0.5	100	1	6

Co: number of the combination;

I-2.2.2 Data Simulations

To study the treatment effects of various scenarios and accrual patterns, we simulated 5000 15-year research course clinical trials for each of the 31 different combinations as described in **Table I-1**. The study was performed using *R* version 3.1.2. We used R library *survival* to conduct the *log-rank* test and *Cox* proportional hazard ratio modeling while R programming language was used to simulate the data.

In each trial, the accrual of the patients was assumed to follow a uniform process over the time period T_{Acc}. The patients were then randomized at a 1:1 ratio into the standard and the experimental groups. When the last enrolled patient had been followed up for a length of time T_{FU}, the analysis was performed. Here, we assumed no patient dropouts and patients who survive the date of analysis were defined as right censored.

The event rate of the standard treatment group (S) of the first trial λ^{S}_{1} is defined by disease scenario (S^{s}_{1} (*1 year*) = 0.5). The event rate in the λ^{E}_{k} experimental treatment group (*E*) is derived from the randomly selected hazard ratio from the historical survival distribution. The event rate of the standard treatment group (*S*) λ^{S}_{k} in the *k*th trial ($k > 1$) of successive trials is determined by the result of the *log-rank* test of the survival distribution between the control and the experimental treatment in the previous trial. After the series of *K* trials at the end of the 15-year period, the overall hazard ratio (HR_O) and probability of detrimental *P*[*HRo* > 1] to assess the performances of the design parameters.

I−3 Results

I−3. 1 Effect of the number of trials and α values on the expected mean overall hazard ratio *E(HRo)*

The expected overall hazard ratio simulated under scenarios 1 and 2 are reported in **Table I-2** and **Figure I-3**. From these tables, we can easily observe the following trend:

First, the expected overall hazard ratio *E(HRo)* decreases monotonically when α value increases from 2.5% to 20% under accrual rates η=100, 200 and 500 patients per year, respectively (scenario 1). This trend is remarkably clear when the trial number is relatively large, which corresponds to a relatively small sample size for each trial. It is worth to note that there was very little improvement on relaxing the *a* value when the number of trials *K* is small, i.e. a relatively large sample size. For example, when accrual rate η=100 patients per year and sample size is 900 patients, the expected overall hazard ratios fluctuate between 0.863 to 0.873 under various α values. Similarly, we also observe

the trend of decreased hazard ratio with increased *a* value for scenario 2. To evaluate the treatment effect of scenario 2, we compare its expected overall hazard ratio with trials in scenario 1 that share the same accrual rate (η=100 patients per year) and number of patients (n =100) (**Error! Reference source not found.**). It was found that the addition of a set-up time to the trial decreased the expected overall hazard ratio, which clearly demonstrates that the more trials that carried out in a given time, the less overall hazard ratio will be expected.

Second, as illustrated above, the expected overall hazard ratio *E(HRo)* decrease monotonically when the number of trials increases for a given α value (**Table I-2** and **Table I-3**). The minimized overall hazard ratio *E(HRo)* for each treatment is identified where the number of trials *K* is largest under accrual rate η=100 and 200 patients per year. When accrual rate increased to a certain degree like η= 500 patients per year in scenario 1 (that also means larger sample size), the minimized overall hazard ratio *E(HRo)* is not always the largest one for all the α values (**Table I-4**). However, the trend is there, i.e. the minimized overall hazard ratio *E(HRo)* located in the area where the number of trials *K* was relatively larger.

Table I-2 Observed mean of the Hazard Ratio *HRo* at 15 years, *E(HRo)* for scenario 1 with an accrual rate η = 100 pts/yr

α Value	*K*, Number of trials (number of patients)									
	1 (1000)	1 (900)	1 (800)	1 (700)	2 (600)	2 (500)	3 (400)	3 (300)	5 (200)	7 (100)
0.025	0.871	0.873	0.883	0.883	0.782	0.806	0.710	0.726	0.632	**0.616**
0.05	0.881	0.864	0.873	0.870	0.763	0.774	0.700	0.710	0.601	**0.592**
0.1	0.870	0.863	0.871	0.865	0.768	0.772	0.676	0.698	0.579	**0.526**
0.15	0.859	0.865	0.864	0.876	0.747	0.757	0.678	0.684	0.571	**0.500**

0.2	0.864	0.867	0.873	0.877	0.742	0.754	0.673	0.676	0.538	**0.479**

*The column in bold denote the lowest value of $E(HR_f)$ in a row, defining the number of trials *K* minimizing *E(HRo)* for each treatment selection criterion.

Table I-3 Observed mean of the Hazard Ratio *HRo* at 15 years, *E(HRo)* for scenario 1 with an accrual rate η = 200 pts/yr

α	K, Number of trials (number of patients)									
Value	1 (1000)	1 (900)	1 (800)	1 (700)	2 (600)	2 (500)	3 (400)	3 (300)	5 (200)	7 (100)
0.025	0.764	0.761	0.693	0.676	0.691	0.643	0.584	0.551	0.549	**0.512**
0.05	0.774	0.770	0.670	0.662	0.686	0.606	0.558	0.497	0.518	**0.486**
0.1	0.756	0.749	0.657	0.660	0.680	0.588	0.534	0.504	0.470	**0.442**
0.15	0.760	0.730	0.665	0.642	0.654	0.580	0.515	0.489	0.465	**0.401**
0.2	0.741	0.751	0.660	0.645	0.672	0.569	0.497	0.479	0.444	**0.359**

*The column in bold denote the lowest value of $E(HR_f)$ in a row, defining the number of trials *K* minimizing *E(HRo)* for each treatment selection criterion.

Table I-4 Observed mean of the Hazard Ratio *HRo* at 15 years, *E(HRo)* for scenario 1 with an accrual rate η = 500 pts/yr

α	K, Number of trials (number of patients)									
Value	1 (1000)	1 (900)	1 (800)	1 (700)	2 (600)	2 (500)	3 (400)	3 (300)	5 (200)	7 (100)
0.025	0.514	0.536	0.537	0.496	0.507	0.464	**0.424**	0.433	0.441	0.500
0.05	0.508	0.524	0.517	0.473	0.476	0.432	0.412	**0.395**	0.425	0.429
0.1	0.502	0.508	0.517	0.447	0.461	0.410	0.389	0.372	0.370	**0.368**
0.15	0.491	0.496	0.490	0.459	0.452	0.397	0.361	0.337	**0.330**	0.333
0.2	0.496	0.482	0.504	0.438	0.437	0.403	0.349	0.337	0.326	**0.312**

*The column in bold denote the lowest value of $E(HR_f)$ in a row, defining the number of trials *K* minimizing *E(HRo)* for each treatment selection criterion.

Table I-5 Observed mean of the Hazard Ratio *HRo* at 15 years, *E(HRo)* for scenario 2 with an accrual rate η = 100 pts/yr

HRo	α Value				
	0.025	0.05	0.1	0.15	0.2

E(HRo)	0.679	0.623	0.592	0.553	**0.536**
*E(HRo)**	0.616	0.592	0.526	0.500	**0.479**

*The column in bold denote the lowest value of $E(HR_f)$ in a row, defining the number of trials *K* minimizing *E(HRo)* for each treatment selection criterion. *E(HRo)** is excerpted data from **Table I-2** that has the same accrual rate (η) and number of patients *n*.

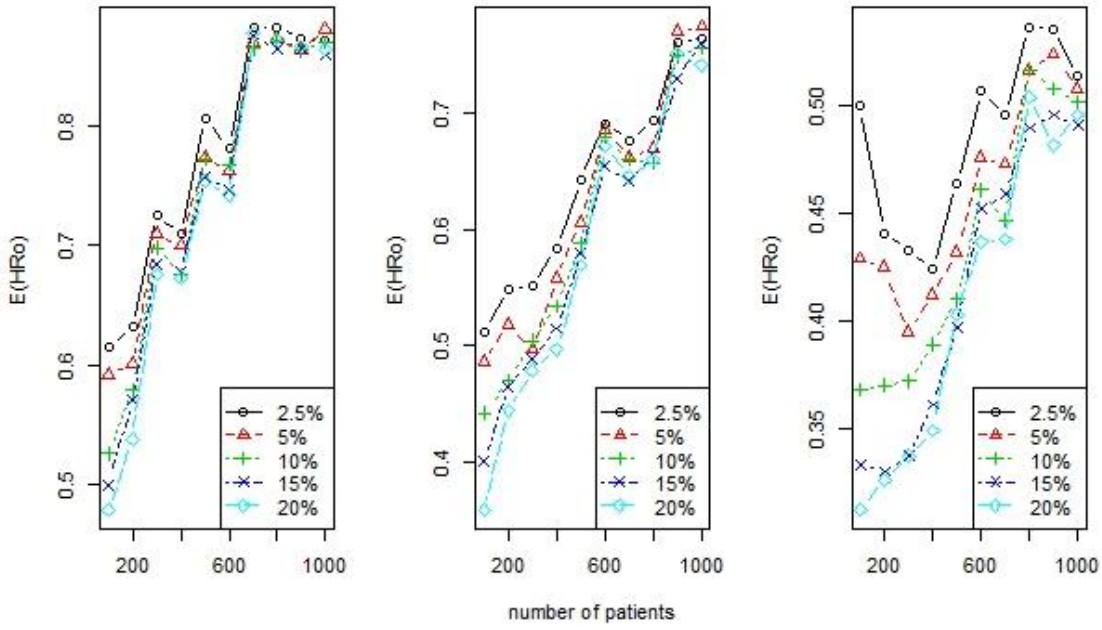

Figure I-3 The expected mean overall hazard ratio *E(HRo)* as functions of the number of patients (number of trials *K*) and α values for the disease scenario 1 (median survival = 1 year, FU = 1 year)

I-3. 2 **Effect of the number of trials and α values on the expected median overall hazard ratio *med(HRo)***

Since the distribution of *HRo* is not normal, we also calculated the median of the overall hazard ratio to summarize the simulated treatment effect. The number of trials and the α value have showed a similar effect on the expected median overall hazard ratio *med(HRo)* (see **Error! Not a valid bookmark self-reference.-Table I-9**, and **Figure I-**

4). The least median overall hazard ratio *med(HRo)* occur where the maximized number of trials *K* were achieved for each specified α value. From a series of simulations with the same number of trials and number of patients at the same accrual rate, we have found that the expected median overall hazard ratio monotonically decreases with larger α levels. These results clearly prove that the smallest trials are associated with the smallest expected hazard ratio, i.e. the best treatment effect.

Table I-6 Observed the median Hazard Ratio *HRo* at 15 years, *Med(HRo)* for scenario 1 with an accrual rate η = 100 pts/yr

α Value	*K*, Number of trials (number of patients)									
	1 (1000)	1 (900)	1 (800)	1 (700)	2 (600)	2 (500)	3 (400)	3 (300)	5 (200)	7 (100)
0.025	1.000	1.000	1.000	1.000	0.854	1.000	0.755	0.784	0.639	**0.595**
0.05	1.000	1.000	1.000	1.000	0.844	0.826	0.735	0.739	0.595	**0.564**
0.1	1.000	1.000	1.000	1.000	0.841	0.828	0.708	0.737	0.580	**0.485**
0.15	1.000	1.000	1.000	1.000	0.811	0.814	0.707	0.722	0.573	**0.466**
0.2	1.000	1.000	1.000	1.000	0.795	0.796	0.718	0.714	0.517	**0.452**

*The column in bold denote the lowest value of *Med(HRf)* in a row, defining the number of trials *K* minimizing *Med(HRo)* for each treatment selection criterion.

Table I-7 Observed the median Hazard Ratio *HRo* at 15 years, *Med(HRo)* for scenario 1 with an accrual rate η = 200 pts/yr

α Value	*K*, Number of trials (number of patients)									
	1 (1000)	1 (900)	1 (800)	1 (700)	2 (600)	2 (500)	3 (400)	3 (300)	5 (200)	7 (100)
0.025	0.823	0.816	0.730	0.718	0.740	0.654	0.569	0.514	0.510	**0.458**
0.05	0.844	0.821	0.699	0.684	0.724	0.619	0.559	0.463	0.481	**0.424**
0.1	0.802	0.793	0.689	0.690	0.719	0.593	0.519	0.474	0.428	**0.379**
0.15	0.811	0.776	0.692	0.668	0.682	0.583	0.507	0.460	0.424	**0.322**
0.2	0.811	0.807	0.687	0.679	0.723	0.576	0.472	0.449	0.415	**0.285**

*The column in bold denote the lowest value of *Med(HRf)* in a row, defining the number of trials *K* minimizing *Med(HRo)* for each treatment selection criterion.

Table I-8 Observed the median Hazard Ratio *HRo* at 15 years, *Med(HRo)* for scenario 1 with an accrual rate η = 500 pts/yr

α	*K*, Number of trials (number of patients)									
Value	1 (1000)	1 (900)	1 (800)	1 (700)	2 (600)	2 (500)	3 (400)	3 (300)	5 (200)	7 (100)
0.025	0.480	0.514	0.530	0.454	0.483	0.426	0.362	0.388	**0.361**	0.434
0.05	0.501	0.514	0.494	0.437	0.436	0.387	0.365	**0.342**	0.351	0.353
0.1	0.486	0.480	0.510	0.420	0.418	0.376	0.339	0.330	0.317	**0.291**
0.15	0.478	0.471	0.476	0.432	0.417	0.354	0.317	0.276	0.280	**0.258**
0.2	0.475	0.469	0.493	0.393	0.404	0.357	0.299	0.271	0.272	**0.254**

*The column in bold denote the lowest value of *Med(HRf)* in a row, defining the number of trials *K* minimizing *Med(HRo)* for each treatment selection criterion.

Table I-9 Observed the median Hazard Ratio *HRo* at 15 years, *Med(HRo)* for scenario 2 with an accrual rate η = 100 pts/yr

HRo	α Value				
	0.025	0.05	0.1	0.15	0.2
Med(HRo)	0.736	0.628	0.583	0.525	**0.506**
Med(HRo)*	0.595	0.564	0.485	0.466	**0.452**

*The column in bold denote the lowest value of *Med(HRf)* in a row, defining the number of trials *K* minimizing *Med(HRo)* for each treatment selection criterion. *Med(HRo)** is excerpted data from **Table I-6** that has the same accrual rate (η) and number of patients *n*.

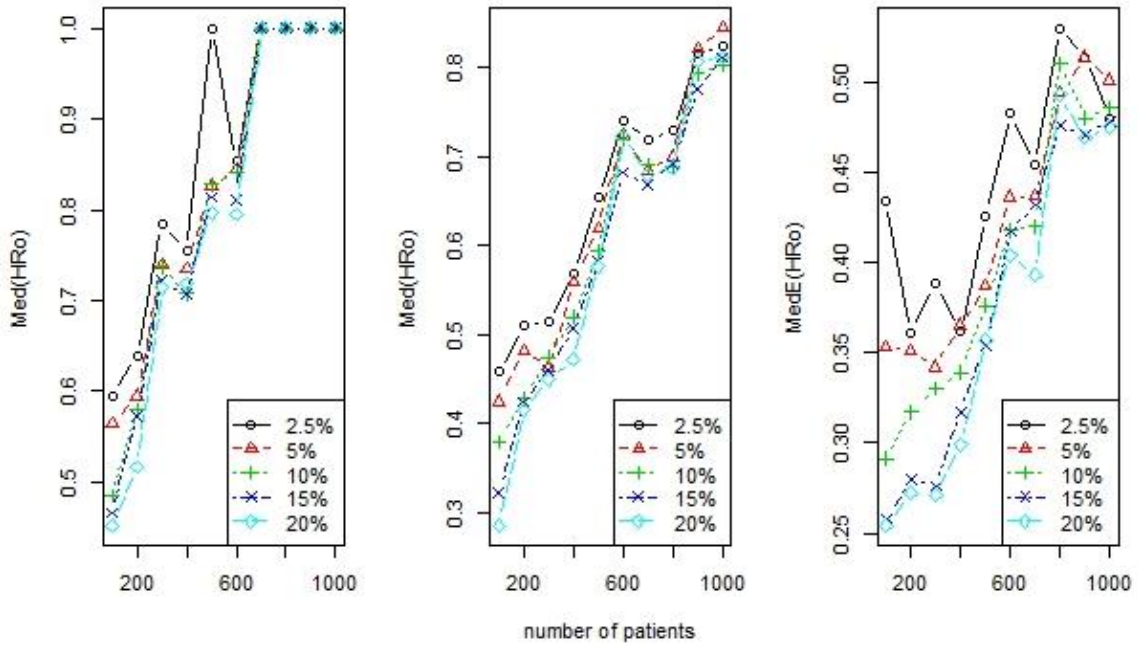

Figure I-4 The expected median overall hazard ratio *Med(HRo)* as functions of the number of patients (number of trials *K*) and *α* values for the disease scenario 1 (median survival = 1 year, FU = 1 year)

I−3. 3 Effect of the number of trials and α values on the probability of a detrimental effect *P(HRO > 1)*

The probability of a detrimental effect *P(HRo > 1)* is the chance that the hazard rate for the new treatment chosen at the end of the 15 years is less effective than the hazard rate for the initial control treatment. The relationship between the probability of a detrimental effect *P(HRo > 1)*, the number of trials and α values is summarized in Table I-10-Table I-13, and Figure I-5. From **Table I-10**, we can easily find that all trials under scenarios 1 and 2 have probability of a detrimental effect less than 1%. It looks like that this happens where the sample size decreases (or *K* increases) or α values turn to bigger. Since the probability of a detrimental effect in all cases is less than our expected value

2.5%, we will not consider this indicator in the following step to find the optimal design parameter.

Table I-10 Probability of a detrimental effect $P(HR_O > 1)$ at 15 years, for scenario 1 with an accrual rate η = 100 pts/yr

α	*K*, Number of trials (number of patients)									
Value	1 (1000)	1 (900)	1 (800)	1 (700)	2 (600)	2 (500)	3 (400)	3 (300)	5 (200)	7 (100)
0.025	0	0	0	0	0	0	0	0	0	0.001
0.05	0	0	0	0	0	0	0	0	0	0
0.1	0	0	0	0	0	0	0	0	0	0.002
0.15	0	0	0	0	0	0	0	0	0	0
0.2	0	0	0	0	0	0	0	0	0	0

Table I-11 Probability of a detrimental effect $P(HR_O > 1)$ at 15 years, for scenario 1 with an accrual rate η = 200 pts/yr

α	*K*, Number of trials (number of patients)									
Value	1 (1000)	1 (900)	1 (800)	1 (700)	2 (600)	2 (500)	3 (400)	3 (300)	5 (200)	7 (100)
0.025	0	0	0	0	0	0	0	0	0	0
0.05	0	0	0	0	0	0	0	0	0	0.002
0.1	0	0	0	0	0	0	0	0	0.001	0.003
0.15	0	0	0	0	0	0	0	0	0	0.002
0.2	0	0	0	0	0	0	0	0.001	0.001	0

Table I-12 Probability of a detrimental effect $P(HR_O > 1)$ at 15 years, for scenario 1 with an accrual rate η = 500 pts/yr

α	*K*, Number of trials (number of patients)									
Value	1 (1000)	1 (900)	1 (800)	1 (700)	2 (600)	2 (500)	3 (400)	3 (300)	5 (200)	7 (100)
0.025	0	0	0	0	0	0	0	0	0	0

0.05	0	0	0	0	0	0	0	0	0	0.002
0.1	0	0	0	0	0	0	0	0	0	0
0.15	0	0	0	0	0	0	0	0	0	0
0.2	0	0	0	0	0	0	0	0	0.001	0.001

Table I-13 Probability of a detrimental effect $P(HR_O > 1)$ at 15 years, for scenario 2 with an accrual rate η = 100 pts/yr

HRo	α Value				
	0.025	0.05	0.1	0.15	0.2
P(HRo > 1)	0.000	0.000	0.001	0.000	0.000
*P(HRo > 1)**	0.001	0	0.002	0	0

*P(HRO > 1)** is excerpted data from **Table I-10** that has the same accrual rate (η) and number of patients *n*.

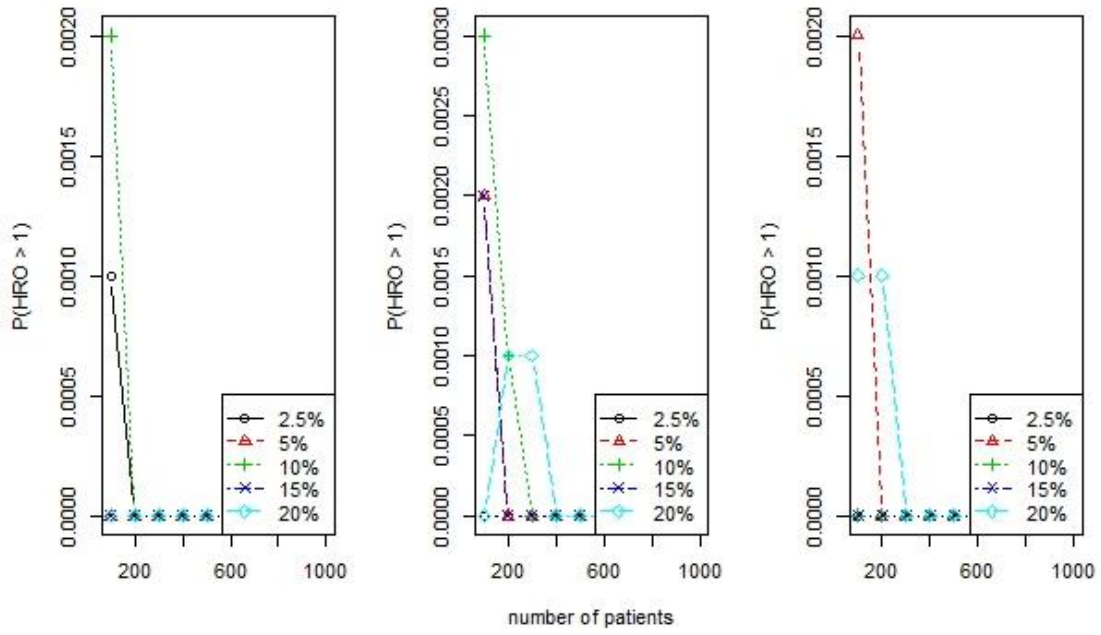

Figure I-5 The probability of detrimental effect *P(HRo>1)* as functions of the number of patients (number of trials *K*) and α values for the disease scenario 1 (median survival = 1 year, *FU* = 1 year)

I-3.4 Search for the optimal design parameter in the different combinations of situations

We have shown in the above section that the treatment effect is related to α values and the number of trials K accomplished in the 15-year time period. The latter determined the trials' sample sizes. In our simulation, we have set up two criteria to optimize the design parameters by considering: 1) minimization of *E(HRo)*; 2) minimization of *E(HRo)* with the constraint $P[HRo > 1] < 0.025$. Since the constraint in the latter criteria is always satisfied in our scenarios, the two rules could be merged as one, i.e. to find the combination of a value and K that yields the lowest expected overall hazard ratio at 15 years.

Let's suppose that there is a clinical trial under the historical distribution having a one year median survival as the baseline and an accrual rate of 100 patients/year, we can find that the optimal design parameters with least mean overall hazard ratio *E(HRo)* 0.479 are the combination of an α values equal to 20% and the number of trials K equal to 7. The corresponding median overall hazard ratio is 0.452 and the probability of a detrimental effect is 0. When the accrual rate is improved to 200 patients/year without changing other settings, the optimal design parameters with least mean overall hazard ratio *E(HRo)* 0.359 are the combination of an α values equal to 20% and the number of trials K equal to 10. The corresponding median overall hazard ratio is 0.285 and the probability of a detrimental effect is 0. When the accrual rate is further increased to 500 patients/year with other settings being the same, the optimal design parameters with least mean overall hazard ratio *E(HRo)* 0.312 are the combination of an α values equal to 20% and the number of trials K equal to 12. The corresponding median overall hazard ratio is

0.254 and the probability of a detrimental effect is 0.1%. It worth to mentioned that some cases exist that the minimized mean overall hazard ratio *E(HRo)* is not associated with the largest number of trial at certain α level when the accrual rate is equal to 500 patients/year. Thus, to get best optimized result, the design parameters should be evaluated for each disease setting and each accrual rate. Our studies have clearly shown that the optimal design parameters by treatment effect is achieved when simulations with the greatest possible number of trials are available. Consider each disease scenario as a whole, we could find that the increase of accrual rate from 100 patients/year to 500 patients/year at the same significance level lead to a substantial decrease the mean overall hazard ratio *E(HRo)* from 0.452 to 0.312 while no obvious change of the probability of a detrimental effect.

I-4 DISCUSSION

We investigated the treatment effect of clinical trials with small sample size by simulation of various combinations of different accrual patterns and α significant levels under historical distribution over a longer research horizon. We evaluated the treatment effect with three metrics: the mean overall hazard ratio *E(HRo),* the median overall hazard ratio *Med(HRo),* and the probability of a detrimental effect *P(HRo>1)*. We have found that the increase of accrual rate from 100 patients/year to 500 patients/year at the same significance level lead to a substantial decrease in the mean overall hazard ratio *E(HRo)* from 0.452 to 0.312 with no obvious change of the probability of a detrimental effect. By relaxing α values from 0.05 to 0.20 while the accrual rate is constant, the mean overall hazard ratio *E(HRo)* decrease monotonically. The optimal design parameters were

chosen based on the minimized mean overall hazard ratio *E(HRo)*. Our calculated results show that smaller sample size, which corresponds to a larger number of trials, and larger α significant levels afford a better treatment effect according to the mean overall hazard ratio *E(HRo)* over a 15-year research period.

Some limitations exist in our study. First, we considered a constant accrual rate during the 15-year trial period. This might not be true and some highly variable accrual patterns might be presented in practice. Thus, the effect of fluctuated accrual rate on the treatment has not been evaluated. Secondly, our simulation relies on the Cox proportional regression model which assumes that a proportional hazard ratio between the control and the experiment groups applied in clinical trials with time-to event outcomes. In reality, the treatment effect might occur under non-proportional hazard ratio assumptions for the control and treatment groups of the study. We have also assumed that a new trial starts just after the previous trial is finished, and that the treatments between the control and the experiment groups are only different on the efficacy without considering the safety profiles of the treatment regimens. Our sensitivity analysis (scenario 2) shows that longer set-up time has some detrimental effect on the overall benefit at the end of 15 years but does not alter our conclusion.

In summary, we showed that accrual pattern and α values play an important role in clinical trials. In order to adapt to our increasingly small patient populations, we suggest that do a multitude of smaller trials with larger alpha values over an extended time period rather than traditional large trials.

Part II Data Analysis of Histone 3 Lysine 36 Trimethylation ChIP-seq in Human Kidney Cancer

II-1 Introduction

Chromatin immunoprecipitation (ChIP) coupled with high-throughput sequencing (ChIP-seq) is a powerful technique for genome-wide determination of the sites of transcription factor binding and histone modification. This chromatin-based assay has several steps (**Figure II-1**).[11] The chromatin preparation step cross-links protein–DNA complexes in living cells with formaldehyde, quenches the reaction with glycine to prevent over-crosslinking, shears the DNA into approximately 100–200 base pair (bp) fragments by sonication or with micrococcal nuclease after cell lysis, and incubates with a specific antibody to pull down the targeted protein of interest (POI). The uncrosslinked and purified DNA fragments are then used for the library preparation as follows. The target-enriched DNA's ends are blunted and added an "A" base to the 3' end, which are ligated wiht an adapter bearing a 3' T-overhang. The ChIP DNA is polymerase chain reaction (PCR) amplified, purified and sequenced on a next-generation sequencing platform.

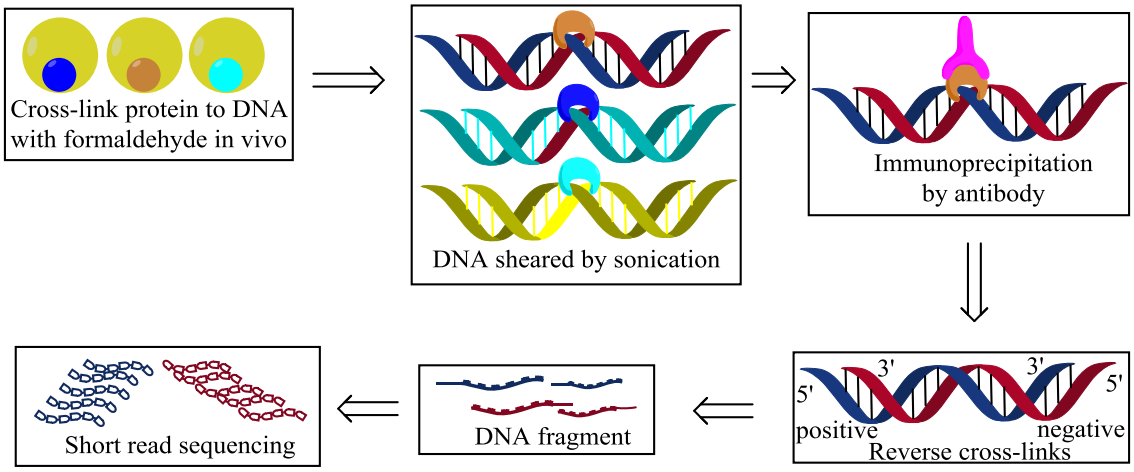

Figure II-1 General procedures for preparation of ChIP library

The raw data processing involves the mapping/alignment of reads to the reference genome, normalization, identification of signal enriched regions (Peak calling), biological interpretation, and visualization of ChIP-seq results (**Figure II-2**). Generally, the majority of ChIP-seq reads are background, whose distribution differs across experiment methods, the genome locations and genome complexity. While peaks with strong signals are biologically most reproducible, it has been found that some modestly-enriched peaks also show high biological regulatory activity.[12] Experiments may also lead to the systematically underrepresented or over-represented signals in the ChIP-seq. To eliminate false positive (non-specific) peaks resulting from chromatin preparation, antibody cross-reaction, PCR amplification, and mapping uncertainties in repetitive regions, a control DNA (input DNA or ChIP DNA from a nonspecific antibody) is usually needed to facilitate the bioinformatics analysis.[13]

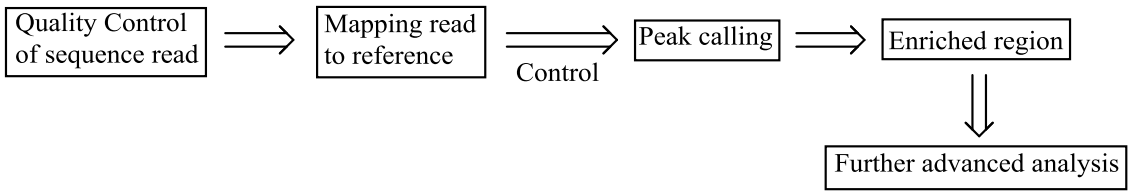

Figure II-2 General workflow for the analysis of ChIP data

There are 3 major categories of profiles based on the range of ChIP-seq tag enrichment. Sharp peaks covering a few hundred base pairs or less, such as the binding sites of transcription factors, are relatively easy to identify and most algorithms available for peak calling are focusing on this class of binding; broad signals from some histone marks (e.g. histone 3 lysine 27 trimethylation ($H_3K_{27}me_3$)) extend over several hundred kilobases.[14] RNA Polymerase II produces a mixture of sharp and broad signals in a length up to a few kilobases.[15] Here, we focus on histone modifications, as these present the most challenging case. To identify high confidence binding sites in a ChIP sample, peak calling algorithms calculating the enrichment of tag density over the background noise are applied. Normalization, which often sets the two samples to have the same total number of uniquely mappable tags, is critical to ensure that the enrichment is not biased toward a sample/region due to systematic errors.

Various studies have revealed that the background portion of the ChIP-seq/the control or different ChIP-seq samples show an approximate linear relationship.[16] Thus most the studies calculated a normalization factor either by a linear scaling of read counts between samples or a linear regression of genomic densities, which is essentially equivalent to the linear scaling of all tags found in samples (**Figure II-3**). In essence, this method involves normalizing the immunoprecipitation (IP) data with the corresponding

control experiment (Input) by a scale factor *s*, which is defined as the ratio of the total number of background tags in IP to that in Input (**Eq. II-1**). The most commonly used strategy to estimate *s* is to divide the reference genome into non-overlapping bins with the width of w and sum the tags in each bin. Then the normalized number of local tags (or each bin) in Input could be obtained by multiplying the corresponding number of raw reads with the scale factor (**Eq. II-2**).

$$s = N_{IP}/N_{Input} = \sum n_{IPb} / \sum n_{Inputb} \qquad \text{Eq. II-1}$$

$$n'_{Input} = n_{Input} \times s \qquad \text{Eq. II-2}$$

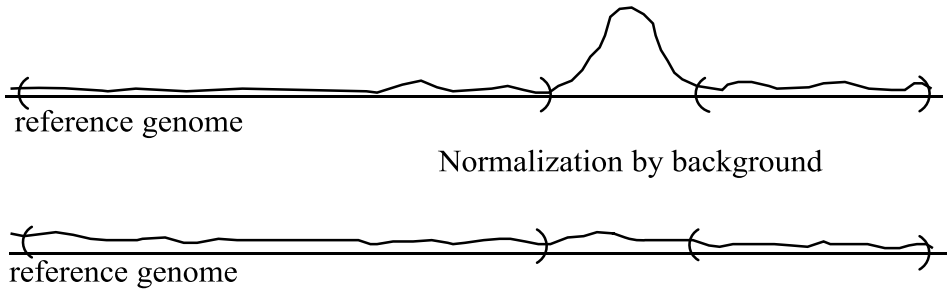

Figure II-3 Normalization by background

Each approach employs a specific set of parameters to estimate the background tags in ChIP-seq. the differences between these methods lie in the choice of bin size and algorithm to determine background tags. Table II-1 summarizes the existing normalization methods.

Table II-1 Commonly used normalization methods

Methods	Normalization parameter	Merit/disadvantage	Ref
Sequencing depth scaling (SDS).	Assuming all tags in ChIP-seq are background and signal in peaks are a small portion of the total tags.	Enriching the background of peaks; untypical peaks are likely to lose information.	17
CisGen	W=100 bp; $n_{IPb} \leqslant 1$	Global parameter estimates;	18

		improved spatial precision; untypical peaks are likely to lose information	
PeakSeq	W=10000 bp; linear regression with a predefined *p*-value	Local statistics; fewer false negatives; sensitive to outliers	19
NCIS	Non-fixed window size; iteration for the first s larger than the previous one	Untypical peaks are likely to lose information	20
MACS	Window size of 1, 5 and 10 kb	Local statistics; fewer false negatives	21

In this study we analyzed $H_3K_{36}me_3$ (deposited by the chromatin regulator $SETD_2$), an active mark associated with transcribed genes.[22]

II-2 Data source

The ChIP-seq raw sequences for $H_3K_{36}me_3$ and input DNA (human kidney cancer) were downloaded from EBI database.[23] The commands and the URLs used for downloading are shown in **Figure II-4**.

```
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR351/ERR351382/ERR351382.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR351/ERR351359/ERR351359.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR351/ERR351366/ERR351366.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR351/ERR351380/ERR351380.fastq.gz
```

Figure II-4 Commands for downloading the sequences

The information about the sequences is listed in the **Table II-2**.

Table II-2 Sequence attribution

Replicate	Immunoprecipitate/input	FASTQ.gz	Total sequence tags
1	$H_3K_{36}me_3$ (IP)	ERR351382	53,826,161
	input	ERR351359	37,459,480
2	$H_3K_{36}me_3$ (IP)	ERR351366	37,903,732

	input	ERR351380	41,734,531

The Input samples for the two individuals are quite different: one produced roughly 16 million fewer sequence tags while another one yield 4 million more sequence tags than the corresponding $H_3K_{36}me_3$ IP samples.

II-3 **Workflow and result**

All the software we used is publicly available. The computational platform is based on the high-performance computing cluster at Minnesota Supercomputing Institute of the University of Minnesota. The majority of the steps in this procedure is done from the command line in the Linux operating system environment.

Our data analysis involves three major steps: 1) Quality checking the sequence to ensure the analyzed tag is in good quality; 2) Map the raw reads back to a reference genome to identify the structure of sequences;[24] 3) Peak calling to identify the differentiate real signal from noise by comparing with the control.[25]

II-3.1 **Quality Control Analysis to Assess Sequence Quality**

To check the sequence run and ensure a good genome alignment, it is important to test the quality of raw sequence data, which will help to identify problems like low-quality base pair or duplicates causing by adapter and primer during library preparation. The open source software FastQC (Babraham Bioinformatics)[26] is one of the widely used software packages for this step. The result for the test is discussed below briefly.

II-3.1.1 **Per Base Sequence Quality**

The boxwhisker plots here show the average (blue line), median (the central red

line) and inter-quartile range (25-75%, the yellow box) of the sequence quality score per base across all reads in a file. From the plots (**Figure II-5**), we could safely conclude that the base quality is pretty good: all bases are in the green region, which indicate very good quality calls. We have also observed that the quality of reads degrades as the run progresses as expected.

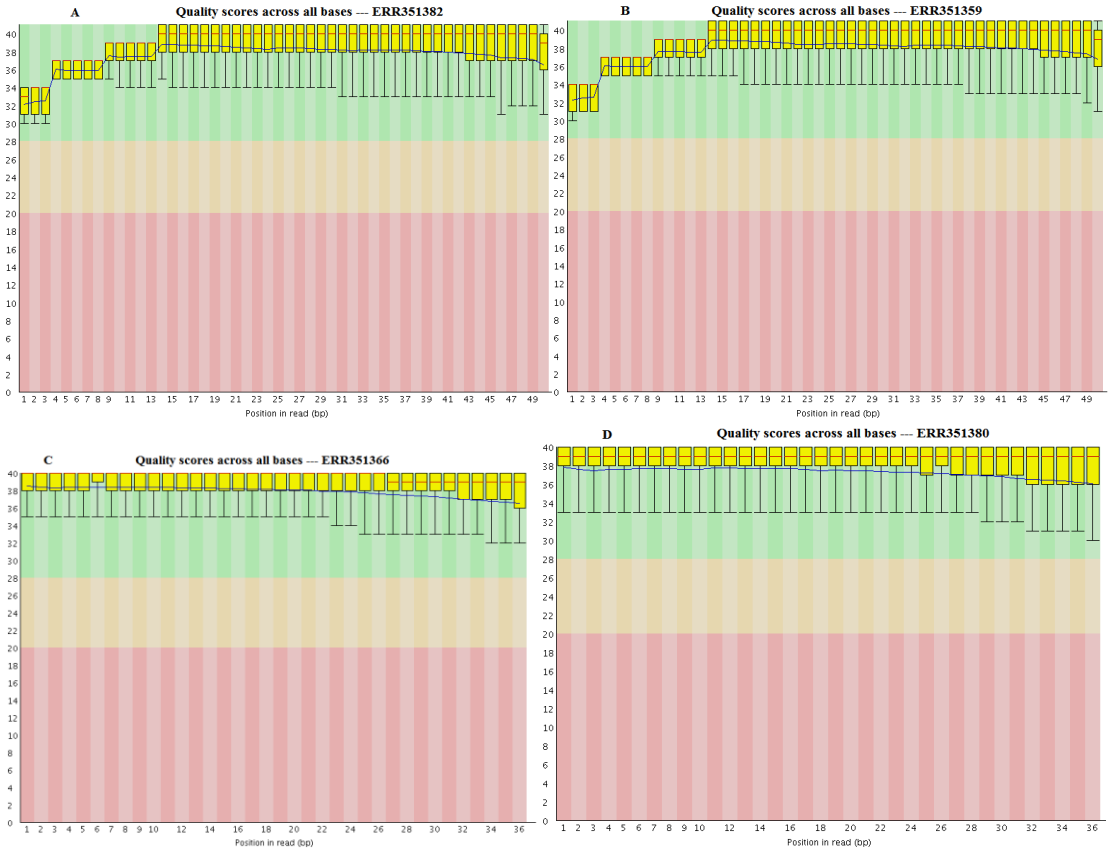

Figure II-5 Per base sequence quality report generated from FastQC analysis.

II-3. 1. 2 **Per Sequence Quality**

The per sequence quality plot is the distribution of the average quality per read, which could be used to determine subset of the reads with low score. Generally, when the average quality per read is below 20 (i.e. 1% error rate), it fails to pass the test. From the

plots (**Figure II-6**), we found that most reads of ChIP samples have an average quality of 38 or 39, suggesting that the vast majority of reads had a good quality.

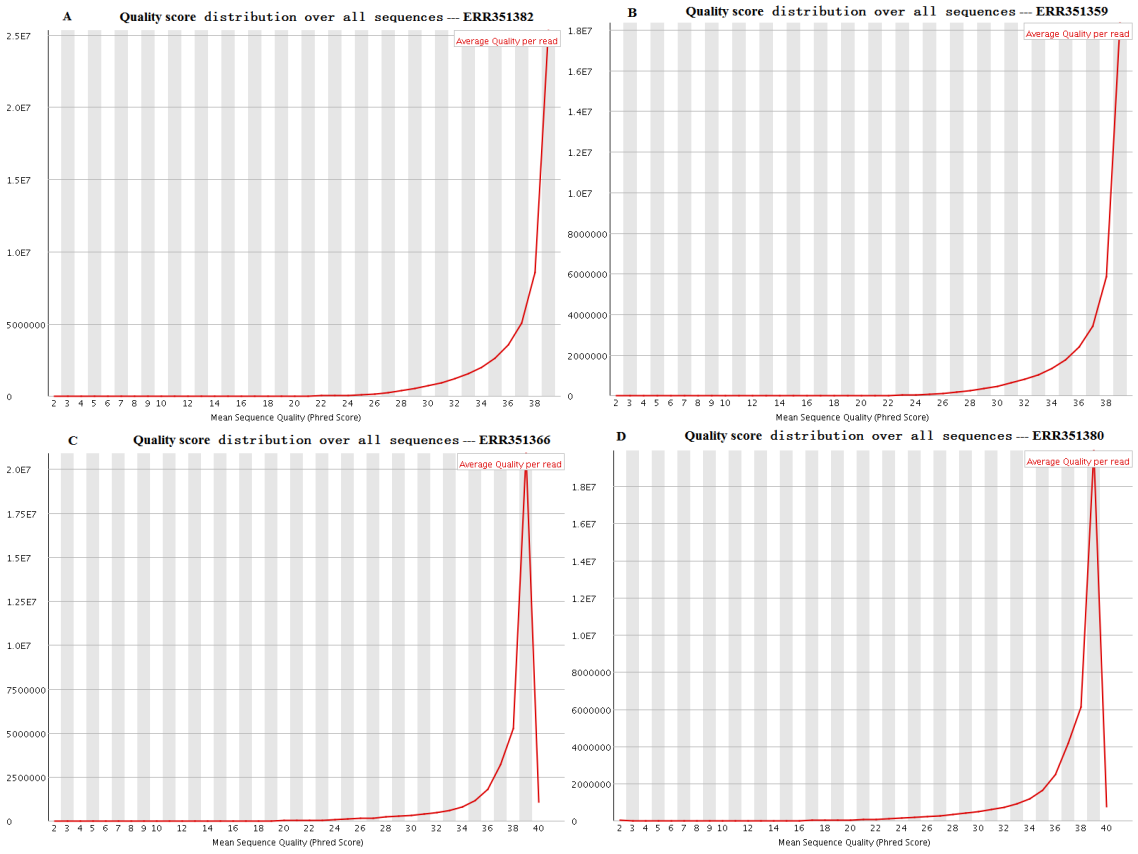

Figure II-6 Per sequence quality report generated from FastQC analysis. Quality score distribution over all sequences for Input (B, D) and $H_3K_{36}me_3$ (A, C).

II-3. 1. 3 **Per Base Sequence Content**

If there is no pattern of special sequences (such as over represented sequence caused by PCR duplicates) presence, the probability of each base (A, T, G, and C) appearing at each position should be the same. Thus the base content across the sequence should be a straight line. In practice, the first few bases usually show some bias, probably due to the presence of not completely random adapters or primers. Usually, the test

determines the difference between the proportion of A and T, or C and G, if the value is greater than 10%, a warning would be given. When the value is over 25%, a failure alert would be issued. Our reports (**Figure II-7**) gave a difference around 10% across the length of the sequence tag except the first a few bases, indicating good quality of the reads.

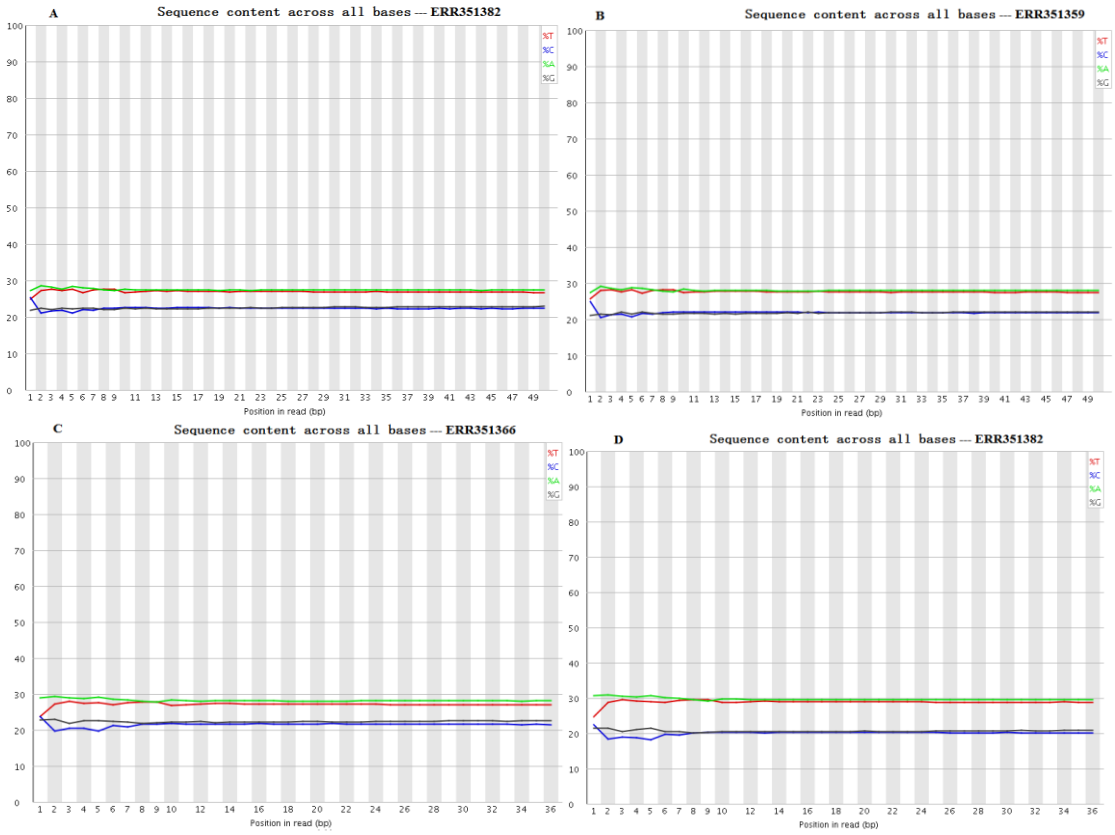

Figure II-7 Per base sequence content report generated from FastQC analysis. Sequence content across all bases for Input (B, D) and $H_3K_{36}me_3$ (A, C).

II-3.1.4 **Per Sequence GC Content**

Different genomic regions have different nucleotide compositions. The GC count per read across sequence reflects the GC content of the corresponding sequenced

genome. In comparing the distribution of the sample to simulated theoretical distribution (its central peak denotes the overall GC content of the genome), if the two distributions deviate from each other, it implies that the library might be contaminated or contain a significantly over-represented sequence. A 15% deviation leads to a warning and over 30% deviation would fail to pass the test. From the reports of the analysis (**Figure II-8**), we found that all four samples passed the test.

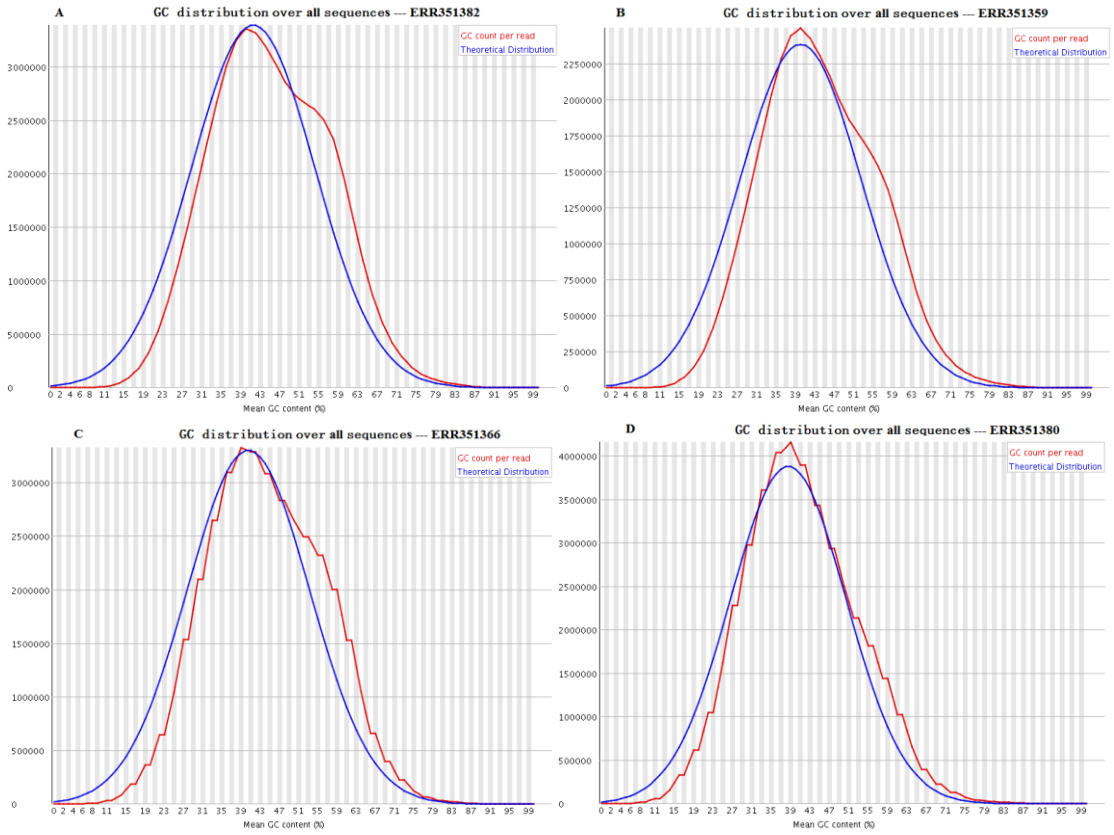

Figure II-8 Per sequence GC content report generated from FastQC analysis. GC content across each sequence for Input (B, D) and $H_3K_{36}me_3$ (A, C).

II-3.1.5 **Per Base N Content**

This test gave the percent of bases that could not be called by the sequencer across

the read. The threshold value for the test to issue a warning or failure alert is 5% or 20%, respectively. All samples passed this quality control test (**Figure II-9** A-D) and only the first two sequences have a few bases that the sequencer could not identify.

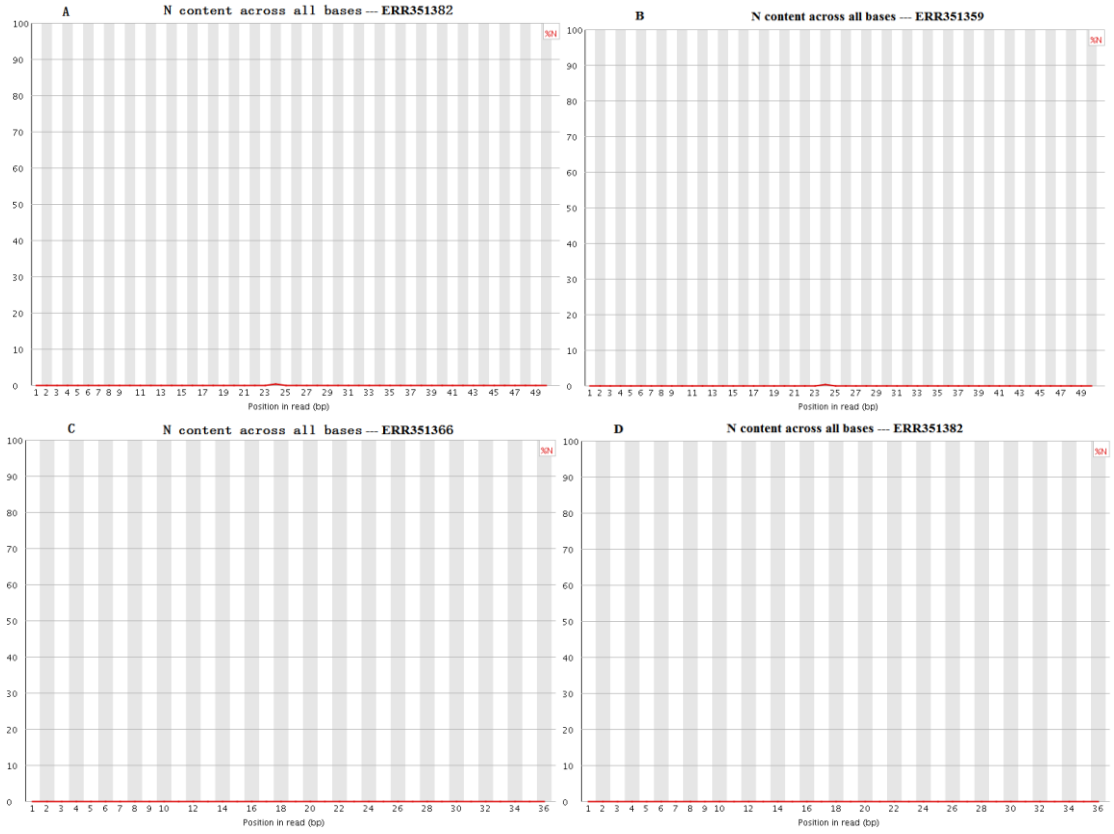

Figure II-9 Per base N content report generated from FastQC analysis. N content across all bases for Input (B, D) and $H_3K_{36}me_3$ (A, C).

II-3.1.6 Sequence Length Distribution

The Distribution of the length of the sequence tags describes the information of read length in each ChIP sample. This test tells issues such as early termination or mispriming during the sequencing run. Our results showed that the two sets of ChIP-seq have a read length at 50 and 36, respectively, and all samples were adequately sequenced

(**Figure II-10** A-D).

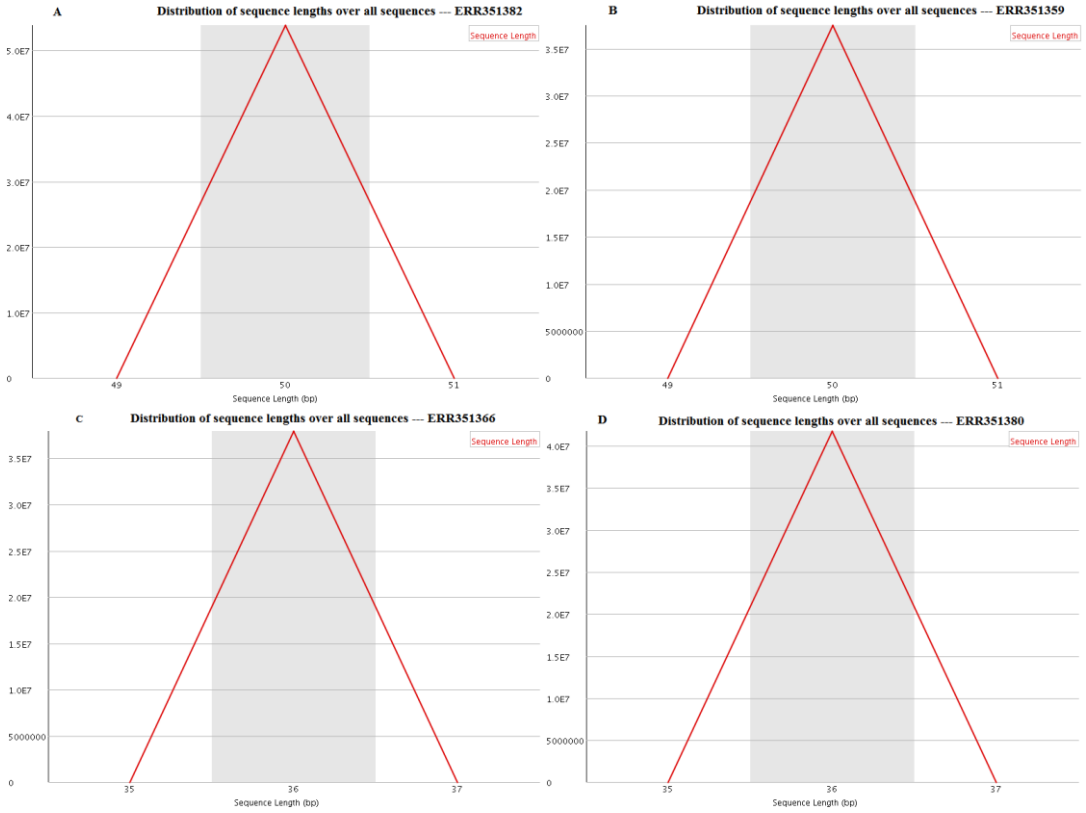

Figure II-10 Sequence length distribution report generated from FastQC analysis. Distribution of sequence lengths over all sequences for Input (B, D) and $H_3K_{36}me_3$ (A, C).

II-3. 1. 7 **Sequence duplication levels**

In order to reduce the computational requirements, this test only considers the first 50 bases of a read across the first 100,000 reads within each file. If the first 50 bases between reads are the same, they will be deemed as identical (duplicates). The blue line on the report shows the distribution of duplication levels of the full 100,000 reads while the red line represents the distribution of the de-duplicated sequences. The duplication levels increased from diverse subsets to specific enrichment of subsets. The test will fail

with over 50% duplicates in the total sequence. Our report shows that all samples have passed this quality check (**Figure II-11**).

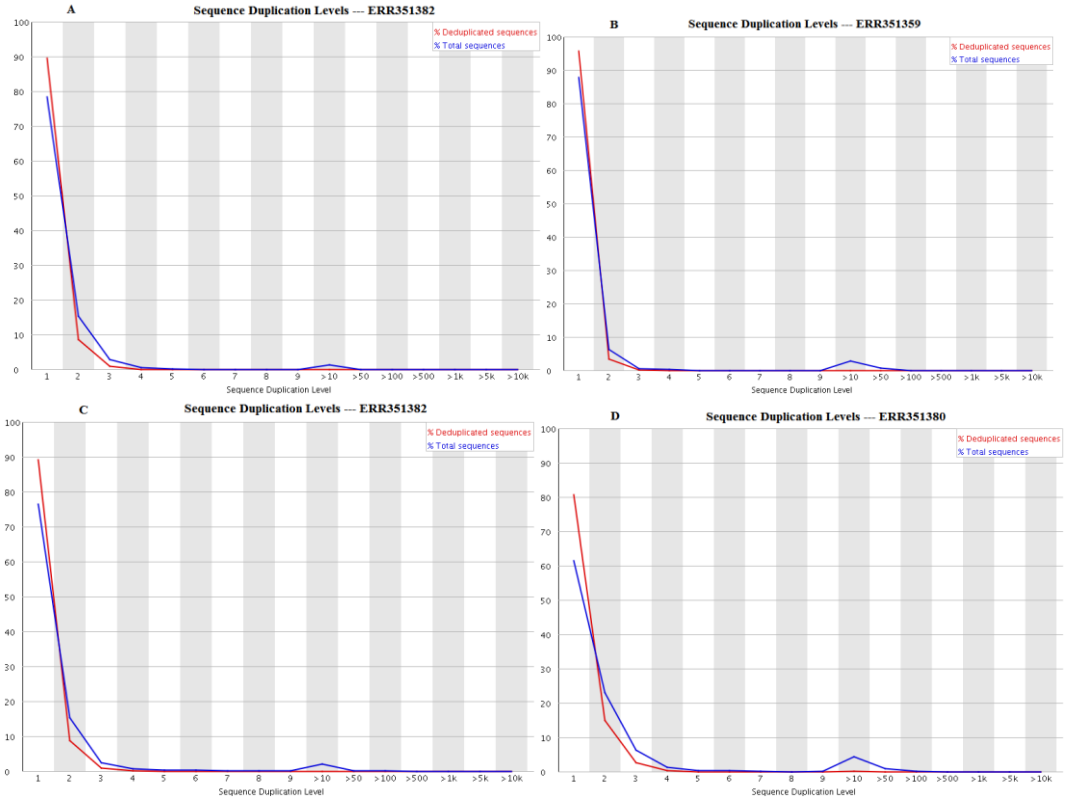

Figure II-11 Duplicate sequence report generated from FastQC analysis. Sequence duplication levels for Input (B, D) and $H_3K_{36}me_3$ (A, C).

II-3.1.8 **Over represented Sequences**

This test gave information about adapter or primer related contamination. It examines the first 200,000 reads in each file and any similar sequences representing over 0.1% of the total reads would be matched with a database of known sequencing primers and adapters. All samples have passed this test and no over represented sequences were observed.

II-3.1.9 **Summary of FastQC analysis**

The result of the FastQC test was summarized in **Table II-3**. All samples passed the check and are good enough for further analysis.

Table II-3 Summary of FastQC analysis

Analysis	Sequence			
	ERR351382	ERR351359	ERR351366	ERR351380
Per Base Sequence Quality	√	√	√	√
Per Sequence Quality	√	√	√	√
Per Base Sequence Content	√	√	√	√
Per Sequence GC Content	√	√	√	√
Per Base N Content	√	√	√	√
Sequence Length Distribution	√	√	√	√
Sequence duplication levels	√	√	√	√
Over represented Sequences	√	√	√	√

II-3.2 **Reads Alignment to the human genome**

After quality control analysis, the next step is to align the short reads from each ChIP-seq and input sample to the human genome. Here we are using the Burrows-Wheeler Alignment (BWA) tool (version 0.5.9), which represents a fast algorithm for the mapping of short-read data. This processing retains reads mapping to unique genomic loci and removes unmappable reads (such as divergent sequences or low-quality sequences) and reads aligning to multiple locations. The output BAM file contains strands, coordinates and other alignment information.

The detailed commands are shown below (**Figure II-12**), where *human.fa* represents the indexed human genome (hg19) generated by “bwa index -a bwtsw”. The command *bwa aln* aligns the reads according to the suffix array (SA) coordinates and produces a binary file *input.sai*, which could only be recognized by *bwa samse*. The

options "*-o, -l, -t, -k, -m*" denote maximum number of gap opens, seed length, number of threads, maximum edit distance in the seed (number of mismatched allowed), and maximum entries in the queue, respectively. The command *samse* is designed for the single-end reads, which transforms the SA coordinated *input.sai* to the chromosomal coordinated *input.sam*. The option *-n* represents the maximum number of alignments to output in the XA tag. The *input.sam* file was then transformed to its binary version (BAM) of Sequence Alignment/Map (SAM) file (smaller in size and faster in manipulation) by Samtools, which consists of a series of utilities that manipulate alignments in the BAM format including sorting, merging and indexing. The *view* command output sequence in the option specified format. The option *-Shb* specifies the input file type to be SAM (*-S*), including header in the output (*-h*) and specifying the output file type to be BAM (*-b*). The flagstat command provides 13 simple statistics like duplicates and mapping ratio on a BAM file. To get a unique read (a single read mapping to one best position), the *view* command was used. The option "*-F 4 -q 20*" extracts the mapped sequences (*-F 4*) with minimum mapping quality 20 (*-q 20*). The result was shown in **Table II-4**, with up to 98.7% reads mapped to the reference.

```
bwa aln -o 1 -l 32 -t 4 -k 2 -m 50000000 indexed_human.fa input.fastq.gz >input.sai
bwa samse -n 10 -f input.sam indexed_human.fa input.sai input.fastq.gz
samtools-0.1.12a/samtools view -Shb input.sam > input.bam
samtools flagstat input.bam
samtools view -b -f 4 input.bam > unmapped.bam
samtools view -b -F 4 file.bam > mapped.bam
samtools view -h -F 4 -q 20 input.bam |samtools view -Sbh - > input.U1.bam
samtools view -c input.U1.bam
```

Figure II-12 Commands for reads alignment

Table II-4 Chip-Seq reads aligned to human genome

Sample	Total raw reads	Mapped reads	% Mapped	Unmapped reads	% Unmapped
ERR351382	53,826,161	52,751,583	98.0%	1,074,578	2.0%
ERR351359	37,459,480	36,308,272	96.9%	1,151,208	3.1%
ERR351366	37,903,732	37,418,837	98.7%	484,895	1.3%
ERR351380	41,734,531	40,990,915	98.2%	743,616	1.8%

Since Chip-seq data involves a PCR amplification process, it inevitably generates duplicates in the final sequences. Research has found that removing duplicate reads, i.e. keeping at most one read per genomic position at each strand, increases the significance of biological signals.[27] Thus in the next step, Picard tools[28] (Java-based command-line utilities) were used to manipulate the bam files. The command *SortSam.jar* sorts the bam file according to chromosome position (SO=coordinate) and the command *MarkDuplicates.jar* removes the duplicates in the bam. The parameters are set as following, and the sorted, deduplicated bam files were produced along with metrics files. The number of reads after duplicates removal are summarized in **Table II-5** by samtool's *view* command with *-c* option. Thus after removing duplicates, we had around 63-79% reads for peak calling.

```
java -Xmx6g -Xms512m -jar picard/SortSam.jar INPUT=input.U1.bam
OUTPUT=input.U1.sorted.bam MAX_RECORDS_IN_RAM=2000000 SO=coordinate
TMP_DIR=working_directory VALIDATION_STRINGENCY=SILENT
java -Xmx6g -Xms512m -jar picard/MarkDuplicates.jar INPUT=input.U1.sorted.bam
```

```
OUTPUT=input.U1.dedup.s1.bam METRICS_FILE=input.U1.sorted_metrics.txt
REMOVE_DUPLICATES=TRUE ASSUME_SORTED=true
VALIDATION_STRINGENCY=SILENT MAX_FILE_HANDLES=100
TMP_DIR=working_directory/input_U1_dedup
samtools view -c *.dedup.s1.bam        #count the number of reads
samtools index *.dedup.s1.bam          #Index the bam files
```

Figure II-13 Commands for removing duplicates

Table II-5 Number of reads after removing duplicates

Sequence	Total sequence tags (fastq.gz)	Mapped reads (U1.bam)	De-duplicated reads (dedup.s1.bam)	Duplicates	Ratio (dedup/total)
ERR351382	53826161	46245451	41007359	5238092	0.76
ERR351359	37459480	31540709	29515907	2024802	0.79
ERR351366	37903732	30064438	26800239	3264199	0.71
ERR351380	41734531	33152881	26187116	6965765	0.63

II-3.3 **Detecting enriched regions**

Once alignments were processed to remove unmapped reads and duplicates, peak calling algorithm was applied to identify binding sites (i.e. peaks). Most peak callers use *p*-value with or without false discovery rate (FDR) to evaluate the confidence of the identified peaks.[29] Here, Sicer (Statistical model for Identification of ChIP-Enriched Regions), specifically designed for the analysis of histone modifications with broad occupancy like $H_3K_{36}me_3$, was employed to find ChIP-enriched regions that are significantly different from a background by comparing ChIP read to a read count threshold derived from a Poisson process (**Figure II-14**).[30] With input as the matched

control, SICER could further improve its accuracy in identifying the "candidate" islands. Since SICER requires BED formatted files as the input, BAM files were converted to BED format in the first step by using the command BamtoBed of the Bedtools suite.[31] The BED format is usually for storing annotations on genomic coordinates. The tab-delimited columns record chromosome name, start position, end position, label, score, and strand. Then the bed files of Chip-seq were subjected to Sicer analysis. The files of IP.U1.dedup.s1.bam.bed and input.U1.dedup.s1.bam.bed correspond to the ChIP and input libraries. The key parameters are redundancy threshold (1 indicating that retaining only 1 read for each set of redundant or duplicate reads), window size (200 bps, default value set for histone modification), fragment size (200 bps for the average size of ChIP fragment, which determines the amount of shift (i.e. 100 bp) from the beginning of a read to the center of the DNA fragment), effective genome fraction (0.75, ratio of the mappable regions to the actual genome size), gap size (600 bps, multiple of window size), and FDR (1E-2, false discovery rate cutoff) for using a control library as background. TMPDIR is the output directory and hg19 specified the species and genome version analyzed. Once the analysis is done, Sicer would produce 10 output files. Of these files, the file *-FDR1E-2 contains detailed information (chromosome, start, end, read count in ChIP library, read count in control library, p-value, fold change, FDR) of identified islands with FDR less than 1% in the ChIP library, the bed file *-W200-G600-FDR1E-2-island.bed is the redundancy-removed raw reads filtered by islands, and the wig file *-W200-G600-FDR1E-2-islandfilterednormalized.wig was used for visualization of the island-fitered ChIP library on a genome browser. The value of fold change reflects

the enrichment of Chip-seq to control (input). Here we identified peaks with fold change great than 2. By applying linux awk command, we first generated a 3-column file *-FDR1E-2_3col.txt with the chromosome, start and end coordinate from the file *-FDR1E-2. Since each line in the *-FDR1E-2_3col.txt file represent an identified peak, we can simply use linux command wc with option -l to obtain the number of peaks. To count the number of reads mapping to the peaks in the Chip-seq, the command intersect in the BEDTools was applied and the corresponding reads overlapped with the peaks was pipelined to samtools' view command. The number of peaks identified and reads intersected with peaks in the Chip-seq were summarized in **Table II-6** and **Table II-7**. We obtain 14,922 and 35,819 peaks respectively for the two sets of sequences.

```
BEDTools/bamToBed -i input.U1.dedup.s1.bam > input.U1.dedup.s1.bam.bed
SICER.sh TMPDIR IP.U1.dedup.s1.bam.bed input.U1.dedup.s1.bam.bed TMPDIR
hg19 1 200 200 0.75 600 1E-2
awk 'BEGIN {FS=OFS="\t"} {if($7>=2) print $1,$2,$3}' *-FDR1E-2>>*-FDR1E-
2_3col.txt
wc -l *-FDR1E-2_3col.txt        # check number of peaks identified
BEDTools/intersectBed -abam IP.U1.dedup.s1.bam -b Peak_file | samtools view -c –
bedtools intersect -abam IP.U1.dedup.s1.bam -b *-FDR1E-2_3col.txt | samtools view –
c - >*_3colpeak_count
head *_3colpeak_count            # check number of peaks identified
```

Figure II-14 Commands for peak calling

Table II-6 Number of identified peaks

Sequence (IP.U1.dedup.s1.bam-W200-G600-islands-summary-FDR1E-2_3col)	Peaks identified	Peaks with fold change >= 2
ERR351382 (IP)	32,391	14,922

ERR351366 (IP)	52,662	35,819

Table II-7 Number of reads in the peaks (Peaks with fold change >= 2)

Sequence	Number of reads
ERR351382 (IP)	9,682,345
ERR351359 (input)	2,806,622
ERR351366 (IP)	6,959,694
ERR351380 (input)	2,375,987

With reads count number in hand, we can easily calculate the ratio of reads mapped to peaks to the total unique reads in each Chip-seq. The ratio is greater than 0.2 for ChIP and around 0.1 for the control (**Table II-8**).

Table II-8 Ratio of reads mapped to peaks to reads count in the Chip-seq

Sequence	Unique Reads in Chip-seq (dedup.s1.bam)	Reads mapped to peak in Chip-seq	Ratio
ERR351382 (IP)	41,007,359	9,682,345	0.236
ERR351359 (input)	29,515,907	2,806,622	0.095
ERR351366 (IP)	26,800,239	6,959,694	0.259
ERR351380 (input)	26,187,116	2,375,987	0.091

II-3.4 Enriched regions through Normalization

Since the read count in the control is unlikely the same as the read count in the background of the Chip-seq library, the raw read counts have to be normalized first before further analysis.[32] Normalization is an essential step to reduce the impact of noise, standardize the representation of data, and facilitate the identification of cluster structure in peak calling. A common method is to use the ratio of total reads or peak reads between

the ChIP and input samples for normalization. This estimation is often accompanied with bias. Here we will use the ratio of mapped reads within non-peak regions (they are true background) for normalization, which should increase the number of peaks and the fold-change of peaks relative to the input.

Before normalization, we draw a scatter plot of the read counts in the Chip-seq library to that in the control library over pre-defined intervals. The purpose of this procedure is to check whether the library-size-based normalization is suitable. To do so, bin-level read count needs to be made. In the first step, the bam files of Chip-seq and input were converted into a tab-delimited 6-column bed file by setting the fragment size at 200 bp. This step was done by linux *awk* command. Then, the number of reads overlapping each of the 500-bp intervals of the human genome was counted. The recorded number of mapped reads (the fourth column) was transformed to a log2 scale to facilitate further manipulation. The detailed procedure is provided in the **Figure II-15**.

```
samtools view *.s1.bam | \
awk 'BEGIN {FS="\t"; OFS="\t"} {if ($2 ==16) print $3,($4+ length($10) - 200),($4-1+ length($10)),".","1","-"; else if ($2 ==0) print $3,$4,($4-1+ 200),".","1","+"}' | \
awk 'BEGIN {FS="\t"; OFS="\t"} {if ($2 <=0) print $1,"1",$3,$4,$5,$6; else print $0}' > *s1.bam.bed
BEDTools /intersectBed -a hg19_24chr.w500.bed -b *s1.bam.bed -bed -wa -c | \
Awk
'{printf("%s\t%d\t%d\t%d\t%.2f\t%.2f\n",$1,$2,$3,$4,log($4+1)/log(2),log(($4*10000000/e5)+1)/log(2))}' > *s1.bam.bed.w500.count.tab     #e5 indicated the actual count in the *.s1.bam file
```

Figure II-15 Commands for calculation normalization variable

With the data in hand, a scatter plot for each of the IP (x-axis) and input (y-axis)

pair was drawn (only Chr1 data was illustrated, **Figure II-16**). From the plots, we observed a large proportion of reads was in the enriched regions in both IPs. This indicates that normalization using peak-less reads should be more appropriate.

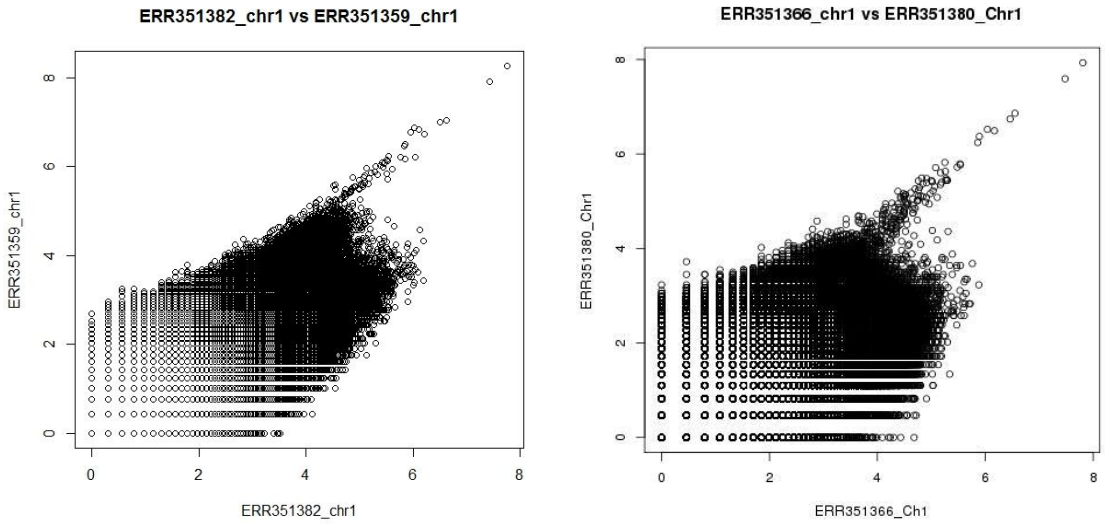

Figure II-16 Scatter plot of IP to its corresponding input

To normalize according to the ratio of mapped reads within non-peak regions, we can calculate the normalization factor based on the following formula (**Eq. II-3**):

NF=(IP_reads_count - IP_reads_in_Peak)/(INPUT_reads_count - INPUT_reads_in_Peak_matching_region) Eq. II-3

Where IP_count and INPUT_count are the number of mapped reads in *.s1.bam; IP_Peak_count and INPUT_Peak_count are the number of mapped reads within peak regions. The result of the calculation was summarized in **Table II-9**. The normalization factors calculated are 1.172 and 0.833 for the replicates RR351382 (IP)/ERR351359 (INPUT) and ERR351366 (IP)/ERR351380 (INPUT) respectively.

Table II-9 Calculation of normalization factor

Data set	IP count	IP Peak count	INPUT count	INPUT Peak count	NF
1	41,007,359	9,682,345	29,515,907	2,806,622	1.172
2	26,800,239	6,959,694	26,187,116	2,375,987	0.833

Data set: 1. ERR351382 (IP), ERR351359 (INPUT); 2. ERR351366 (IP), ERR351380 (INPUT)

With the normalization factor NF in hand, we can adjust the fold change in the peak files and calculate the number of peaks after the normalization accordingly. The commands we used are shown in the **Figure II-17** and the result is summarized in **Table II-10**. It showed that the number of reads in the peak increased by 30.2% for sequence ERR351366 and 46.0% for ERR351382.

```
awk 'BEGIN {FS=OFs="\t"} {print $0,$4/($5*NF)}' *islands-summary-FDR1E-2 >
islands-summary-FDR1E-2.re.txt
awk 'BEGIN {FS=OFs="\t"} {if ($7 >=2) print $0}' islands-summary-FDR1E-2 |wc –
l    #peaks before normalization
awk 'BEGIN {FS=OFs="\t"} {if ($9 >=2) print $0}' islands-summary-FDR1E-2.re.txt
|wc -l
```

Figure II-17 Commands for normalizing the peak counts

Table II-10 Enrichment after normalization

Library	IP Peak count before normalization	IP Peak count after normalization	Peak count change
ERR351382	14,922	21,788	46.0%
ERR351366	35,819	46,620	30.2%

II-4 **Conclusion**

Chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) experiment is an important approach to determine the sites of protein-DNA interactions. It has been used to map transcription factor binding, RNA Polymerase II occupancy, and sites of chromatin regulators and histone modifications across cell types, tissues and conditions. To compare a ChIP sample with a control or another ChIP sample, the data has to be normalized to remove experimental biases. A widely used strategy involves linear scaling of number of uniquely mapped reads, which may not be well suitable for cases where the two libraries have very different enrichment levels.

In this research, we have developed a general workflow for the analysis of publicly available histone 3 lysine 36 trimethylation ($H_3K_{36}me_3$) data from human kidney cancer. By estimating the normalization factor only from background, we demonstrated the better performance of our method over the existing approach. An increase of peak counts of 30%% and 46% were observed from the replicates we used. This approach is helpful for identifying reads that were otherwise ignored in the traditional SDS method.

BIBLIOGRAPHY

[1] a) Friedman, L. M., Furberg, C. D., DeMets, D. L., *Fundamentals of Clinical Trials*, **2010**, Springer, New York; b) Pocock, S. J., *Clinical Trials: A Practical Approach*, **2013**, John Wiley & Sons, New York

[2] a) Freiman, J. A., Chalmers, T. C., Sknowledge mith, H., J., *The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial*, The New England Journal of Medicine, **1978**, 299, pp. 690-694; b) Zhang, X., and Long, Q., *Stochastic Modeling and Prediction for Accrual in Clinical Trials*, Statistics in Medicine, **2010**, 29, pp. 649-658.

[3] Lachin, J. M., and Foulkes, M. A., *Evaluation of Sample Size and Power for Analyses of Survival with Allowance for Nonuniform Patient Entry, Losses to Follow-Up, Noncompliance, and Stratification*, Biometrics, **1986**, 42, pp. 507-519.

[4] a) Kleinbaum, D. G., Klein, M., *Survival analysis: a self-learning approach* (2nd ed.) Springer, New York, NY **2005**; b) Spruance, S. L., Reid, J. E., Grace, M., Samore, M., *Hazard ratio in clinical trials Antimicrob Agents Chemother*, **2004**, 48, pp. 2782–2792; c) Hernán, M., *The hazards of hazard ratios Epidemiology*, **2010**, 21, pp. 13–15; d) Zhang, D., and Quan, H., *Power and Sample Size Calculation for Log-Rank Test with a Time Lag in Treatment Effect*, Statistics in Medicine, **2009**, 28, pp. 864-879.

[5] a) Hale M, Gillespie W.R., Gupta S., Tuk B., Holford, N. H., *Clinical trial simulation. Streamlining your drug development process*, Appl. Clin. Trials, **1995**, 5, pp. 35–40; b) Johnson, S.C.D., *The role of simulation in the management of research: what can the pharmaceutical industry learn from the aerospace industry?*, Drug Inf, J., **1998** 32, pp. 961–969;

[6] Holford, N., Ma, S. C., Ploeger, B. A., *Clinical Trial Simulation: A Review*, Clinical Pharmacology and Therapeutics, **2010**, 88, pp. 166-182.

[7] Strauss, N. and Simon, R., *Investigating a sequence of randomized phase II trials to discover promising treatments*, Statistics in Medicine, **1995**, 14, pp, 1479-1489.

[8] Sposto, R., Stram, D.O., *A strategic view of randomized trial design in low-incidence paediatric cancer*, Stat. Med., 1**999**, 18, pp. 1183–1197.

[9] a) Marie-Cécile Le D., Karla V B., Julien M., Daniel S., *Taking the long view: how to design a series of Phase III trials to maximize cumulative therapeutic benefit*, Clin. Trials, **2012**, 9, pp. 283-292; b) Bayar M. A., Le Teuff G, Michiels S, Sargent D., Le Deley M. C., *New insights into the evaluation of randomized controlled trials for rare diseases over a long-term research horizon: a simulation study*, Stat Med., **2016**, 35, pp. 3245-4358.

[10] Djulbegovic B., Kumar A., Soares H. P., et al., *Treatment success in cancer: New cancer treatment successes iden-tified in phase 3 randomized controlled trials conducted by the National Cancer Institute-sponsored cooperative oncology groups, 1955 to 2006*, Arch. Intern. Med., **2008**, 168, pp. 632–642.

[11] a) Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Zhang, J., *Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data*, *PLoS Computational Biology*, **2013**, 9, e1003326; b) Bardet A.F., He Q., Zeitlinger J., Stark A., *A computational pipeline for comparative ChIP-seq analyses*, Nat. Protoc., **2012**, 7, pp. 45–61; c) Yan H., Evans J., Kalmbach M., et al., *HiChIP: a high-throughput pipeline for integrative analysis of ChIP-Seq data*, BMC Bioinformatics, **2014**, 15, pp. 280-291; d) Thomas-Chollier, M., et al. *A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs*, Nat. Protoc., **2012**, 7, pp. 1551–1568.

[12] Johnson D. S., Mortazavi A., Myers R. M., *Genome-wide mapping of in vivo protein-DNA interactions*, Wold B Science, **2007,** 316, pp. 1497-502.

[13] a) Teytelman L., Ozaydin B., Zill O., Lefrançois P., Snyder M., Rine J., Eisen M. B., *Impact of chromatstructures on DNA processing for genomic analyses*, PLoS One, **2009**, 4, e6700; b) Ambrosini G., Dreos R., Kumar S., Bucher P., *The ChIP-Seq tools and web server: a resource for analyzing ChIP-seq and other types of genomic data*, BMC Genomics, **2016**, 17, pp. 938.

[14] a) Cavalcante, R. G., Lee, C., Welch, R. P., Patil, S., Weymouth, T., Scott, L. J., & Sartor, M. A., *Broad-Enrich: functional interpretation of large sets of broad genomic regions*, Bioinformatics, **2014**, 30, pp. i393–i400.

[15] Baugh L. R., Demodena J., Sternberg P.W., *RNA Pol II accumulates at promoters of growth genes during developmental arrest*, Science, **2009**, 324, pp. 92–94.

[16] a) Rozowsky J., Euskirchen G., Auerbach R., Zhang Z., Gibson T., Bjornson R., Carriero N., Snyder M., Gerstein M., *PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls*, Nat Biotechnol., **2009**, 27, pp. 66–75; b) Zheng,W., Zhao,H., Mancera,E., Steinmetz,L., Snyder, M., *Genetic analysis of variation intranscription factor bindingin yeast*, Nature, **2010**, 464, pp. 1187–1191; c) Zhong, M., Niu, W., Lu, Z., Sarov, M., Murray, J., Janette, J., Raha, D., Sheaff er, K., Lam, H., Preston, E., et al., G*enome-wide identification ofb inding sites defines distinct functions for Caenorhabditis elegans PHA-4/FOXA in development and environmental response*, PLoS Genet, **2010**, 6, e1000848.

[17] Strom, C.Y., Vega, A., et al, *Discovery of estrogen receptor alpha target genes and response elements in breast tumor cells*, Genome Biology, **2004**, 5, R66.

[18] Ji H., Jiang H., Ma W., Johnson D., Myers R., Wong H., *An integrated software system for analyzing ChIP-ChIP and ChIP-seq data*, Nat Biotechnol., **2008**, 26, pp. 1293–1300.

[19] Rozowsky J, et al., P*eakseq enables systematic scoring of ChIP-seq experiments relative to controls*, Nat. Biotechnol., **2009**, 27, pp. 66-75.

[20] Liang K., Keleş S., *Normalization of ChIP-seq data with control*, BMC Bioinformatics, **2012**, 13, pp. 199.

[21] Zhang, Y. et al., *Model-based analysis of ChIP-seq (MACS)*, Genome Biol., **2008**, 9, R137.

[22] a) Carvalho S., Raposo A. C., Martins F. B., Grosso A. R., Sridhara S.C., Rino J., Carmo-Fonseca M., de Almeida S. F., *Histone methyltransferase SETD2 coordinates FACT recruitment with nucleosome dynamics during transcription*, Nucleic Acids Res., **2013**, 41, pp. 2881–2893; b) Radeepa M. M., Sutherland H. G., Ule J., Grimes G. R., Bickmore W. A., *Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing*, PLoS Genet **2012**, 8, e1002717.

[23] http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1936/samples/?s_page=1&s_pagesize=500

[24] Li, H., Durbin, R., *Fast and accurate short read alignment with Burrows–Wheeler transform*, Bioinformatics, **2009**, 25, pp. 1754-1760.

[25] a) Farnham P. J., *Insights from genomic profiling of transcription factors*, Nat. Rev. Genet., **2009**, 10, pp. 605-616; b) Pepke S., Wold B., Mortazavi, A., *Computation for ChIP-seq and RNA-seq studies*, Nat. Methods, **2009**, 6, pp. S22-S32.

[26] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[27] a) Leleu M., Lefebvre G., Rougemont J., *Processing and analyzing ChIP-seq data: from short reads to regulatory interactions. Briefings in Functional*, Genomics, **2010**, 9, pp. 466-476; b) Dozmorov M. G., Adrianto I., Giles C. B., et al. *Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data*, BMC Bioinformatics, **2015**,16, pp. S10.

[28] http://broadinstitute.github.io/picard/

[29] a) Benjamini Y., Hochberg Y., *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, J. Roy. Stat. Soc. B. Met., **1995**, 57, 2pp. 89-300; b) Bailey T., Krajewski P., Ladunga I., et al., *Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data*, PLoS Computational Biology, **2013**, 9, pp. e1003326.

[30] http://home.gwu.edu/~wpeng/Software.htm

[31] http://bedtools.readthedocs.org/en/latest/

[32] a) Liang K., Keles S., *Normalization of ChIP-seq data with control*, BMC Bioinformatics, **2012**, 13, pp. 199-208; b) Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., et al., *The DNA-encoded nucleosome organization of a eukaryotic genome*, Nature **2009**, 458, pp. 362-366.