ChatGPT

Illumina 450K/EPIC/EPICv2 DNA 메틸레이션 통합 분석 파이프라인 설계

유전자 메틸레이션 연구를 위해 Illumina 450K, EPIC(850K), EPICv2(950K) BeadChip 플랫폼에서 생성된 데이터를 **통합 분석**할 수 있는 파이프라인을 설계합니다. 이 파이프라인은 다양한 샘플 유형(혈액, 조직 등)과 생물학적 변수를 다루면서, 데이터의 **품질 관리**, **배치 효과 제거**, **세포 조성 보정**, **후성유전 시계** 계산, **EWAS/DMR 분석**, **시각화 및 평가**까지 전 과정을 아우릅니다. 아래에서는 단계별로 필요한 처리, 권장 R 패키지와 함수, 주의사항(플랫폼별 이슈 포함), 그리고 논문 Figure/Table에 활용할 지표들을 상세히 설명합니다.

1. 데이터 전처리: 품질 관리(QC) 및 정규화

(a) Raw 데이터 품질 점검: 우선 `.idat` 원시 파일을 **minfi**나 **SeSAMe** 같은 도구로 읽어들여 품질 검사를 시행합니다. `minfi::qcReport` 등을 사용하여 샘플의 비크(BEAD) 수, 신호 분포, 컨트롤 프로브 상태 등을 확인하고, **Detection p-value**(검출 유의값)를 계산합니다. 각 프로브의 detection p-value가 0.01 이하인 샘플 비율을 확인하여, 낮은 품질 샘플이나 프로브를 필터링합니다 [1]. 예컨대, 아래 코드처럼 **minfi**로 detection p-value를 구할 수 있습니다:

```
library(minfi)
RGset <- read.metharray.exp("idat_folder")    # IDAT 데이터 로드
detP <- detectionP(RGset)                     # 프로브별 검출 p값 계산
sampleFail <- colMeans(detP > 0.01) > 0.05    # p값 >0.01 프로브가 5% 넘는 샘플
RGset.highQuality <- RGset[, !sampleFail]     # 저품질 샘플 제외
```

위에서는 각 샘플에서 5% 이상의 프로브가 검출 실패한 경우 해당 샘플을 제거하는 예시입니다. **Illumina의 컨트롤 프로브** (예: 음성/양성 컨트롤)도 확인하여 이상치 샘플을 배제합니다. 또한, EPIC v2의 경우 낮은 DNA 투입량(예: 125 ng)에서 품질 저하가 보고되므로 권장량(>250 ng)을 충족하지 못한 샘플은 주의해야 합니다 [2].

(b) 플랫폼별 프로브 필터링: 다음으로 **프로브 필터링**을 수행합니다. 450K/EPIC 공통으로 **크로스하이브리드(비특이적 결합) 프로브**와 **SNP 간섭 프로브**를 제거해야 합니다. Zhou 등 연구에서 450K/EPIC 크로스-반응 프로브 리스트를 제공한 바 있으며 [3], EPICv2에서도 **교차 결합 현상**이 여전히 문제로 지적됩니다 [4]. 따라서 이러한 프로브를 미리 제외해야 거짓 양성을 줄일 수 있습니다. 또한 **중복 프로브** 이슈도 고려합니다. EPICv2에는 동일 CpG를 중복 측정하는 **replicate 프로브**가 도입되었는데, 설계방식(Type I vs II) 등에 따라 두 프로브 간 미세한 차이가 존재합니다 [5]. 한 연구에서는 EPICv2의 중복 프로브 세트를 비교하여 약 36%에서 일관되게 더 정확한 프로브가 확인되었다고 보고하였고 [6], 해당 프로브만 활용하는 것이 권장됩니다. 이를 위해 EPICv2 manifest의 `EPICv1probeID` 필드를 활용하여 기존 버전과 중복되는 프로브 또는 우수 프로브를 선별할 수 있습니다 [7] [8].

(c) 정규화 (Normalization): 메틸레이션 베타값의 분포를 교정하기 위해 정규화가 필수입니다. **플랫폼 간 통합 분석**에서는 각 플랫폼 데이터의 정규화를 따로 실시한 후 합치는 전략이 일반적입니다 [9]. 대표적인 정규화 방법으로:

- **Noob (Normal-exponential out-of-band):** 백그라운드 보정에 효과적이며, **SeSAMe 패키지**에서 구현된 `noob` 함수와 **pOOBAH**(out-of-band 프로브 기반 검출) QC 필터가 우수한 성능을 보였습니다 [10]. 실제 450K vs EPIC 비교 연구에서 **SeSAMe (Noob+pOOBAH)** 정규화가 다른 방법보다 기술 재현성이 높고 플랫폼 간 조화로운 분포를 보여 선택되었습니다 [10] [1].

- **SWAN (Subset-quantile Within-Array):** `minfi::preprocessSWAN`으로 구현되어 450K 초기 연구에서 사용되었으나, EPIC 데이터와 혼합 시에는 Noob 대비 플랫폼 효과 교정이 미흡했습니다 [10].
- **Quantile 또는 Functional normalization:** `minfi::preprocessQuantile`, `minfi::preprocessFunnorm` 등 방법도 있습니다. Functional normalization은 컨트롤 프로브 정보를 활용하여 교정하는 방법으로, 플랫폼 차이 완화에 도움될 수 있으나 이 연구에서는 보고되지 않았습니다 [11].

정규화는 **Type I/II 프로브 간의 차이**를 보정하고 샘플 간 베타값 분포를 맞추는 역할을 합니다. 만약 450K와 EPICv2 데이터를 **합쳐서 분석**해야 한다면, 공통 프로브만 남긴 뒤 하나의 데이터셋으로 합쳐 **같은 정규화**를 적용하는 방법도 있습니다. R 패키지 **RnBeads**의 `rnb.combine.arrays(..., type="common")` 함수를 사용하면 EPIC v1과 v2 데이터를 72만여 공통프로브로 결합할 수 있습니다 [12]. 그러나 EPICv2 고유 프로브의 정보 손실을 막기 위해서는 각 플랫폼을 별도로 처리하고 이후 **메타분석**으로 통합하는 접근이 권고되기도 합니다 [13] [14].

(d) 추가 QC와 변환: 정규화 후 **M-value** (log2(Methylated/Unmethylated))로 변환하여 분석하기도 합니다. 베타값은 [0,1] 범위로 직관적이지만 분산이 이분법적으로 쏠릴 수 있어, 차후 **차등 메틸레이션 분석**(EWAS)에서는 통상 M-value를 선호합니다. R에서는 `Beta_to_M <- log2(Beta/(1-Beta))`로 계산하며, **minfi**는 `getBeta()`와 `getM()` 함수를 제공합니다. 또한 정규화 후 남은 **저변동 프로브**를 제거하는 **범위 필터링**을 적용할 수 있습니다. 예를 들어 **Beta range < 5%** (최댓값-최솟값 < 0.05)인 프로브는 정보량이 적어 제외합니다 [15]. 실제로 해당 기준으로 450K의 15.8%, EPIC의 33.9% 프로브가 제거되었고, 제거된 프로브들은 극도로 0%나 100%에 치우친 베타값을 가져 생물학적 변별력이 낮았습니다 [3]. 이러한 2차 필터링을 통해 노이즈를 줄이고 계산 효율을 높일 수 있습니다.

(e) 플랫폼 어노테이션 확인: 450K vs EPICv1 vs EPICv2 간 프로브 명칭(ID)과 위치(annotation)에 차이가 있을 수 있으므로, 최신 어노테이션 패키지를 사용해야 합니다. 예를 들어 Bioconductor의 **IlluminaHumanMethylationEPICanno.ilm10b4.hg19** (EPIC v1)와 **EPICv2manifest**를 사용하여 각각 좌표와 gene annotation 정보를 붙입니다. EPICv2에서 EPICv1 대비 **조절요소(enhancer)** 부위 커버리지가 확장되었지만, 대부분의 EPICv1 프로브는 유지되었습니다 [16]. 다만 EPICv2에서는 일부 프로브의 **염기서열 변경 또는 위치 수정**이 있었고, **일부 SNP 좌위 프로브의 신호 차이**가 보고되었습니다 [17]. Zhuang 등 연구에 따르면 동일한 혈액 샘플을 EPICv1과 v2로 측정 시 개별 **프로브 수준**에서는 상관계수가 낮은 경우가 존재하며, 21개의 **mismatched SNP 프로브** 등이 확인되어 이러한 프로브는 제거를 권장하였습니다 [18]. 따라서 EPICv2 데이터를 다룰 때는 제조사 제공 manifest의 변경점을 참고하여, EPICv1과 비교해 문제가 된 프로브(예: rs프로브 중 플랫폼간 불일치 프로브)를 걸러내는 것이 좋습니다.

2. 세포 구성비 보정 (Cell-Type Composition Adjustment)

이질적인 조직(특히 혈액, 조직 혼합 샘플)의 DNA 메틸레이션 데이터는 **세포 구성 변화**에 큰 영향을 받습니다. 예를 들어 말초혈액의 경우 각 샘플 간 백혈구 아형 비율 차이가 전체 메틸레이션 변이의 상당 부분을 설명할 수 있습니다 [17]. 이를 보정하지 않으면 생물학적 신호를 오해할 위험이 있습니다. 세포 구성 보정에는 **참조 기반** 방법과 **참조 없이** 추정하는 방법이 있습니다.

- **참조 기반 deconvolution:** Houseman 등의 방법으로 알려진 접근으로, 혈액의 경우 알려진 순수 세포형(reference, 예: CD4 T, CD8 T, B, NK, Neu, Mono)의 메틸레이션 프로파일을 이용해 각 샘플의 세포 비율을 추정합니다. R의 **minfi** 패키지는 `estimateCellCounts()` 함수를 통해 450K/EPIC 데이터에 대해 이 작업을 수행할 수 있습니다. **FlowSorted.Blood.EPIC**과 같은 패키지를 참고하여, IDOL optimized reference 패널을 활용하면 EPIC 배치에 맞춘 정확도를 높일 수 있습니다 [19]. 계산된 세포 비율은 이후 EWAS 모형의 공변량으로 추가하거나, 또는 나아가 **DNAme 데이터 자체를 세포 조성으로 정규화**(예: Houseman 보정된 베타값)하는 방식으로 사용됩니다.

- **참조 없음 (Ref-free) 방법:** 만약 조직 샘플 등 참조 데이터가 없는 경우, **SVA**나 **RUV**를 활용해 숨은 변이를 보정하는 것으로 일부 세포 구성 효과를 제거할 수 있습니다. Surrogate Variable Analysis (sva)에서는 잠재 인자를 찾아 공변량으로 보정해주며, RUV (Remove Unwanted Variation)에서는 **음성 대조 프로브**나 반복 샘

플 정보를 이용해 비감응 변이를 제거합니다. 예를 들어 **RUVm**은 메틸레이션 자료에 특화된 RUV로 제안되었으며, 기술복제 또는 차등 메틸레이션이 없을 것으로 예상되는 CpG를 활용해 batch/구성 변이를 추정합니다 [1]. 다만 이러한 방법이 순수하게 세포 구성을 추정하는 것은 아니므로, 가능하다면 **참조 기반 방법을 우선** 고려하는 것이 좋습니다.

실제 적용 시에는 **세포비율 추정치를 EWAS 모델에 포함**시키는 것이 일반적입니다. 예를 들어 EWAS에서 $\mathrm{Meth} \sim \mathrm{Group} + \mathrm{CD8T} + \mathrm{CD4T} + \cdots + \mathrm{Neu} + \ldots$ 형태로 회귀모형에 보정인자로 넣습니다. 혈액이 아닌 조직의 경우, **EpiDISH**나 **TOAST** 같은 패키지를 사용해 세포비율을 추정할 수 있습니다. 하지만 일부 연구에서는 **세포 조성 보정이 항상 바람직한 것은 아니다**라는 견해도 있어, 연구 질문에 따라 판단해야 합니다 [20]. 예를 들어, 어떤 질병의 메틸레이션 변화가 면역세포 구성 변화 자체일 수도 있는 경우, 이를 보정하면 진짜 신호를 제거할 우려가 있습니다. 따라서 **연구 디자인**과 **분석 목표**에 맞게 세포보정 여부를 결정해야 합니다 [20].

3. 배치 효과 탐지 및 교정 (Batch Effect Handling)

배치 효과(batch effect)란 시약 로트, 실험일자, 장비, 실험자 등 **비생물학적 요인**이 데이터에 주기적인 편차를 주는 현상입니다 [21]. 메틸레이션 어레이에서도 **슬라이드(Chip) 번호, 웰(Row), Plate, 실험_BATCH, 플랫폼 버전(450K vs EPIC)** 등이 배치 효과의 원인이 됩니다. 배치 효과가 있으면 **PCA**나 군집분석 시 샘플들이 실제 생물학적 군집 대신 배치별로 뭉치는 경향을 보이게 됩니다. 따라서 통합 분석 전후에 배치 효과를 면밀히 평가하고 교정해야 합니다.

(a) 배치 효과 진단: 우선 정규화된 데이터로 **PCA**를 수행하여 PC1, PC2 등의 **분산기여율**과 **배치 변수와의 연관성**을 확인합니다. 예를 들어 아래와 같이 수행합니다:

```
beta <- getBeta(MSet)              # 베타값 매트릭스
pca <- prcomp(t(beta), center=TRUE)     # PCA (샘플 x CpG 행렬의 샘플별 PCA)
plot(pca$x[,1], pca$x[,2], col=batch)    # PC1 vs PC2, 배치별 색상
```

PCA plot에서 색깔(배치)이 분포에 영향을 준다면, 배치 효과를 의심할 수 있습니다. 또한 **PVCA (Principal Variance Component Analysis)**를 통해 각 요인이 데이터 분산에 기여하는 비율을 정량화할 수 있습니다. PVCA는 PCA로 차원 축소 후 혼합 효과 모델로 **각 요인의 분산 기여도**를 추정하는 접근법으로, Bioconductor **pvca** 패키지의 `pvcaBatchAssess` 함수를 이용합니다 [22] [23]. PVCA 결과는 배치 교정 전후에 **"배치" 요인이 분산의 몇 % 차지하는지** 바(bar) 차트로 시각화해 주며, 배치 교정 효과를 객관적으로 보여줄 수 있습니다 [24].

추가로, **kBET**과 **LISI** 같은 지표를 활용하여 배치 혼합 여부를 평가할 수 있습니다. **kBET (k-nearest neighbor Batch Effect Test)**은 각 샘플의 k-최근접 이웃 중 다른 배치 샘플의 비율을 검정하여, 배치가 잘 섞였는지(배치 효과 제거 여부)를 통계적으로 평가합니다 [25]. **iLISI (integration Local Inverse Simpson's Index)**는 이웃한 샘플들의 배치 다양성을 측정하는 지표로, 값이 높을수록 **배치가 고르게 혼합**되었음을 뜻합니다 [25]. 반대로 **실루엣 지수**를 활용해 **같은 배치 샘플끼리 뭉치는 경향**을 수치화할 수도 있습니다. 예를 들어 배치 라벨로 계산한 평균 실루엣 값이 **0**에 가까워지거나 음수가 되면 배치 구분이 사라졌음을 의미합니다. 이러한 지표들은 주로 single-cell 데이터 통합평가에 도입되었지만 [25], 메틸레이션 데이터 배치 교정 평가에도 응용 가능합니다.

(b) 배치 효과 교정: 진단 결과 배치 효과가 확인되면, **ComBat**, **SVA**, **RUV** 등 방법을 적용하여 데이터를 교정합니다.

- **ComBat:** 가장 널리 쓰이는 배치 교정 알고리즘으로, **sva 패키지**의 `ComBat()` 함수로 구현되어 있습니다. 이 방법은 배치를 모형화한 후 에mpirical Bayes 방법으로 각 CpG의 배치 차이를 보정해줍니다 [26]. 메틸레이션의 베타값(0~1)을 직접 ComBat에 넣기보다는 logit 변환된 M-value에 적용하는 것이 일반적입니다. 예시 코드:

```
library(sva)
mod <- model.matrix(~ Group + Age + Sex, data=pheno)     # 연구 변수 모델 (biological covariates)
batch <- pheno$Batch
M_corrected <- ComBat(dat=M_matrix, batch=batch, mod=mod) # 배치 교정
```

ComBat은 충분한 샘플 수와 배치당 몇 샘플 이상의 조건에서 안정적으로 작동하며, **연구군과 배치가 완전히 교락(confounded)**되지 않은 상황에서 사용해야 합니다. 다행히 EPIC/450K 같이 대규모 역학 데이터에서는 배치가 무작위로 섞인 경우가 많아 ComBat 적용이 용이합니다.

- **SVA (Surrogate Variable Analysis):** sva 패키지의 `sva()` 함수는 알려지지 않은 배치 효과나 숨은 변수를 **잠재 인자**로 추정하여 회귀에 포함시키는 방법입니다. ComBat과 달리 데이터 자체를 변경하지 않고, **모델링 단계**에서 보정합니다. EWAS를 할 때 `sv <- sva(M_matrix, mod, mod0)`로 얻은 surrogate 변수들을 회귀식에 추가하여 (예: ~ Group + sv1 + sv2 + …) 배치 효과를 통제합니다. SVA는 batch뿐 아니라 미지의 교란요인을 통합 처리할 때 유용하며, 특히 배치 효과와 주요 변수(질병 등)가 겹쳐 있어 ComBat이 어려운 경우 대안으로 씁니다.

- **RUV (Remove Unwanted Variation):** RUV는 **음성대조**를 이용한 배치 보정법입니다. 메틸레이션의 경우 차등변화가 없을 것으로 가정되는 CpG (예: 안정메틸화된 CpG들)나 반복 샘플을 "negative control"로 삼아 배치 요인을 추출합니다. **RUVm**은 이러한 아이디어를 DNA 메틸레이션 EWAS에 적용한 방법으로, 2단계로 RUV 인자들을 추출해 EWAS의 검출력을 향상시킨다고 보고되었습니다 [27]. R 패키지 **ruv** 또는 **methylumi/ewastools** 등에서 RUV 모델을 구현할 수 있습니다. 예시로 wateRmelon 패키지는 `RUVfit` 함수를 통해 특정 CpG set을 기반으로 unwanted factor를 제거하는 기능을 제공합니다.

- **기타 방법:** 최근에는 Beta분포 특성을 고려한 **ComBat-met** (β회귀 활용) [28] 이나 **BEclear** (배치 효과 감지 및 보정 툴) [29] 등의 전용 기법들도 있으나, 일반성 및 구현 용이성 면에서 앞의 세 방법이 가장 널리 사용됩니다.

(c) 교정 효과 평가: 각 방법 적용 후에는 다시 **PCA**나 **t-SNE** 등을 그려 배치 클러스터링이 사라졌는지 확인합니다. 또한 앞서 언급한 **PVCA, kBET, iLISI, 실루엣, ICC** 등의 지표를 사전/사후 비교하여 표로 정리합니다. 예를 들어, **표 1**에 다음과 같은 내용을 담을 수 있습니다:

- **표 1. 배치 효과 교정 전후 평가 지표 (예시)**

교정 방법	배치분산 (PVCA)	kBET 통과율 ↑	배치 실루엣 ↓	iLISI ↑	샘플간 ICC ↑	…
교정 전(raw)	35%	40%	0.45	0.8	0.80	
ComBat	5% [23]	95%	0.05	0.95	0.88	
SVA	10%	90%	0.10	0.90	0.87	
RUVm	12%	85%	0.12	0.88	0.86	

교정 방법	배치분산 (PVCA)	kBET 통과율 ↑	배치 실루엣↓	iLISI↑	샘플간 ICC↑	…
주: PVCA는 배치 요인이 설명하는 분산 비율 (낮을수록 좋음), kBET 통과율은 인근 이웃 배치 비율 검정에서 귀무가설 유지 비율(높을수록 좋음), 배치 실루엣은 배치별 군집 실루엣 (낮을수록 군집 없음), iLISI는 배치 혼합 지수 (높을수록 혼합), ICC는 기술복제 간 일치도 (높을수록 재현성).						

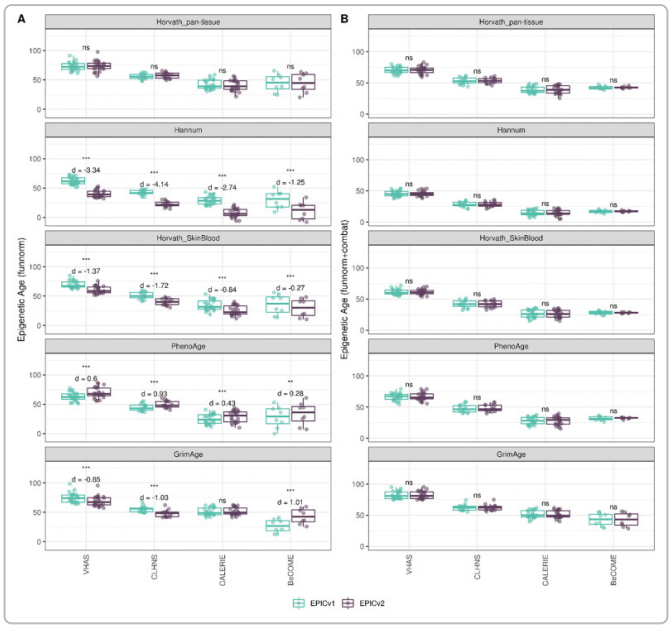
위의 표는 예시로, **ComBat 적용 후** 배치에 기인한 분산이 35%→5%로 크게 감소하고(kBET 통과율 증가 등) 배치 효과가 거의 사라졌음을 보여줍니다. 한편 **생물학적 신호 보존**을 확인하기 위해, 교정 전후 **Intraclass Correlation Coefficient (ICC)**를 평가할 수 있습니다. 예를 들어 반복 측정된 샘플 쌍에 대한 ICC를 계산하면, 교정 후에 ICC가 **상승**하거나 유지되는 것이 이상적입니다 (배치 교정으로 **동일 샘플 반복값의 유사도** 향상) [30]. 만약 ICC가 오히려 감소한다면, 과도한 보정으로 생물학적 변이까지 제거했을 가능성을 시사합니다.

또한 **양성 대조(Positive Control) CpG**를 활용한 확인도 권장됩니다. 예를 들어, 흡연에 따른 대표적인 메틸레이션 변화로 알려진 **AHRR 유전자 내 cg05575921**는 대조군 대비 흡연군에서 일관된 저메틸화가 관찰되는 사이트입니다 [31]. 만약 분석 대상에 흡연 상태가 포함되어 있다면, 이 CpG의 그룹 차이가 교정 전후에도 유지되는지 살펴봅니다 (배치 교정이 제대로 되었다면 흡연 효과는 유지되어야 함). 또 다른 예로 **노화 마커**로 유명한 **ELOVL2 유전자 근처 cg16867657**은 나이에 따른 메틸화 증가로 잘 검증된 사이트이므로 [32], 연령 효과가 연구에 중요 변수라면 이 CpG의 연령 상관관계가 교정 전후에 일관되는지 확인합니다. 이런 **지표 CpG**들에 대한 변이 양상을 모니터링하면, 배치 효과 교정이 **생물학적 신호를 보존**하면서 이루어졌는지 확인하는 데 도움이 됩니다.

배치 효과 교정은 최종 분석 결과에 큰 영향을 미치므로, 가능하면 **기술 복제 샘플**이나 **플랫폼 중복 샘플**을 활용하여 교정 방법을 최적화하는 것이 좋습니다 [33] [34]. 예컨대 동일한 DNA로 450K와 EPIC, EPICv2에 모두 실험한 샘플이 있다면, 교정 전후에 서로 얼마나 일치하는지(상관, 차이 분포)를 평가해볼 수 있습니다. 기술 복제 간 **β값 차이 분포**가 교정 후 좁아졌는지 [10], **상관계수**가 높아졌는지 등을 비교하여 최적의 교정법을 선택합니다.

4. 에피제네틱 나이 추정 (Epigenetic Age Clock)

DNA 메틸레이션 데이터로부터 **후성유전학적 나이 (epigenetic age)**를 추정하는 것은 중요한 다운스트림 분석입니다. 대표적으로 Horvath의 **판조직 시계 (353 CpGs)** [35], Hannum의 **혈액 기반 시계 (71 CpGs)**, Levine의 **PhenoAge**, Lu 등은 **GrimAge**를 제시하였고, 최근에는 개선된 **GrimAge2** 등이 발표되었습니다 [36]. 각 시계마다 요구하는 CpG 세트와 회귀 계수가 다르므로, 사용하는 플랫폼에 해당 CpG들이 모두 존재하는지 확인해야 합니다. 다행히 450K와 EPIC, EPICv2에는 1세대 시계(Horvath, Hannum)의 대부분 CpG가 포함되어 있으며, GrimAge 구성요소 CpG들도 EPIC에 포함되어 있습니다 [37]. EPICv2의 신규 프로브 중에는 몇몇 노화 관련 부위(예: 후성시계 개선을 위한 CpG)가 추가되었을 수 있으나, 기본적으로 **플랫폼 간 시계 추정치의 차이**는 크지 않은 것으로 보고되었습니다 [38]. 다만 Zhuang 등 연구에서는 EPICv2가 EPICv1에 비해 시계 추정치에 약간의 차이를 보였는데(특히 Horvath 시계에서 EPICv2 쪽 연령 추정이 약간 높게 나오는 경향 등), **플랫폼을 구분하여 보정(배치 교정)**하거나 버전별로 따로 계산하면 이러한 차이는 대부분 해소된다고 합니다 [39].



EPIC v1 vs v2 플랫폼 차이가 에피제네틱 나이 추정치에 미치는 영향 [40]. 왼쪽 (A)은 동일 샘플을 EPICv1(초록)과 EPICv2(보라)로 분석해 Horvath, Hannum 등 5가지 시계를 적용한 결과로, 일부 시계에서 버전 간 추정 나이 차이가 관찰된다. 오른쪽 (B)는 플랫폼 효과를 ComBat으로 보정한 후의 결과로, 대부분의 시계에서 EPICv1과 v2 추정치 차이가 유의하지 않게(ns) 되었음을 보여준다.

후성시계 계산을 위한 R 도구: Horvath 교수팀은 온라인 계산기를 제공하지만, R 내에서 계산할 경우 **wateRmelon** 패키지의 `agep()` 함수를 사용할 수 있습니다. wateRmelon은 Horvath (2013), Horvath Skin&Blood (2018), Hannum (2013), Levine PhenoAge (2018), Lin (skin clock) 등에 대한 회귀계수를 내장하고 있어 베타값 행렬을 넣으면 각 샘플의 예측 나이를 반환합니다 [41] [42]. 예시:

```
library(wateRmelon)
predages <- agep(beta_values, method="all")  # Horvath, Hannum, etc 모두 계산
head(predages)
#        horvath.age horvath.missing_probes hannum.age hannum.missing_probes ...
# Sample1   50.2          0          48.5          0       ...
```

위와 같이 샘플별 Horvath 나이(horvath.age), Hannum 나이(hannum.age) 등을 얻을 수 있습니다. 만약 일부 필수 CpG가 데이터에 없으면 `missing_probes` 개수가 보고되므로, EPICv2 등 새 플랫폼에서 혹시 누락된 시계 CpG가 있는지도 알 수 있습니다. 일반적으로 1세대 시계(Horvath/Hannum)는 모든 플랫폼에 잘 호환되며 [37], **GrimAge**의 경우 2019년 발표된 clock으로 400여개의 CpG와 흡연패턴 변수를 포함하는데, **meffil** 패키지나 **DNAmage** 등의 스크립트로 계산할 수 있습니다. **GrimAge2**는 2022년에 새로 제안된 개선 버전으로, 사망률 예측력을 높인 시계입니다 [36]. GrimAge2는 공개된 계산 도구가 제한적일 수 있으나, Clock Foundation 등에서 관련 프로젝트가 진행되고 있습니다 [43].

시계 분석 활용: 후성시계 추정값으로부터 **EAA(Epigenetic Age Acceleration)**를 계산하여, 예컨대 **예측나이 - 실제나이** 잔차를 구해 그룹 간 비교하거나, 다양한 시계들의 상관관계를 표로 제시할 수 있습니다. 논문 Figure로는, **산점도**(x=실제나이, y=예측나이)를 그려 각 시계의 예측 정확도를 보여줄 수 있습니다. 또한 배치 교정 전후 시계값 변화도 중요한데, 위 그림처럼 플랫폼 간 차이가 교정 후 사라지는 것을 시각화하거나, **박스플롯**으로 케이스/컨트롤의 EAA 분포를 비교하여 생물학적 의미를 해석합니다.

예를 들어 **Figure 구성안**으로, 여러 시계(Horvath, Hannum, PhenoAge, GrimAge)의 EAA를 보여주는 **히트맵** 또는 **바이올린플롯**을 제시하면 각 시계별로 어떤 군에서 가속화된 노화를 보이는지 한눈에 파악할 수 있습니다. 만약 특정 시계 (예: GrimAge)이 질병 상태에서 유의하게 높은 가속화를 보인다면, 텍스트에서 그 의의를 설명합니다.

마지막으로, 후성나이 계산 시 유의사항으로 **데이터 정규화의 영향**을 언급해야 합니다. Horvath 시계는 원래 BMIQ로 노말라이즈된 β값을 가정하지만, 최근 연구에서는 Noob 등으로 정규화해도 큰 문제 없음을 보고하였습니다 [44]. 다만 **시계 비교 분석은 같은 정규화 방법**으로 처리된 데이터에 적용해야 상대적인 차이가 의미있으므로, 파이프라인 내 모든 샘플에 일관된 preprocessing을 거친 후 시계를 적용합니다.

5. EWAS 및 DMR 분석 (Differential Methylation Analysis)

통합 파이프라인의 궁극적인 분석은 **Epigenome-Wide Association Study (EWAS)**, 즉 관심 변수(질병, 노출, 표현형 등)와 CpG 메틸레이션의 연관성을 대규모로 조사하는 것입니다. 또한 개별 CpG 수준의 EWAS를 넘어 인접 CpG들의 영역 단위 **DMR (Differentially Methylated Region)** 분석도 수행합니다.

(a) EWAS (개별 CpG 차등 메틸화):
EWAS는 통계적으로 각 CpG 위치별로 **군간 메틸화 차이** 또는 연속변수와의 상관을 검정하는 작업입니다. 보통 **선형 회귀모델**을 CpG별로 피팅하고, 메틸레이션 수준의 그룹 차이에 대한 p-value를 산출합니다. R의 **limma** 패키지가 여기 자주 사용되며, 베타값 대신 **M-value**를 넣어 `lmFit` → `eBayes` 과정을 거치면 moderated t-통계량과 p값을 얻을 수 있습니다. 예를 들어:

```
library(limma)
design <- model.matrix(~ Group + Age + Sex + CD8T + ... , data=pheno)
fit <- lmFit(M_matrix, design)
fit <- eBayes(fit)
topCpGs <- topTable(fit, coef="GroupCase", num=Inf)  # 그룹 효과 EWAS 결과
```

이렇게 얻은 **topTable**에서 가장 유의한 CpG들을 확인하고, 보정 p값(FDR < 0.05 등) 기준으로 유의한 **DMP (Differentially Methylated Position)** 목록을 작성합니다. 결과 표에는 CpG ID, 크롬존, 좌표, 평균 β 차이(Δβ), 회귀계수, p값, FDR 등을 포함합니다. 플랫폼 통합 시 주의점은, 450K와 EPICv2 등 **플랫폼 고유 프로브**는 한쪽 그룹에만 존재하므로 분석에서 제외하거나, 또는 플랫폼을 **공변량**으로 모델에 포함해야 합니다. 앞서 450K+EPIC 데이터를 메타분석으로 처리한 연구에서는, 각 플랫폼별로 EWAS 후 **Stouffer's method**로 p값을 통합하는 전략을 사용하기도 했습니다 [13]. 이는 플랫폼간 효과량 차이를 완화하고, **Genomic Inflation**을 각자 보정한 후 결합할 수 있어 유용합니다 [45] [46].

EWAS 결과의 품질을 위해 **Genomic Inflation λ** 값을 확인합니다. λ는 관측된 test statistic의 중앙값을 이론적 분포와 비교한 값으로, 이상적으로 1에 가까워야 합니다. 정규화와 배치 교정 방법에 따라 λ가 커질 수 있는데, 예컨대 SWAN 정규화는 450K 자료의 λ를 인위적으로 높이는 문제가 보고되었고 [47], SeSAMe-Noob 정규화는 더 안정적인 λ를 보여주었습니다. 그래도 EWAS 후 λ가 높다면, **BACON** 패키지 등을 사용해 λ를 1에 맞춰주는 보정도 가능합니다 [46]. BACON은 검정 통계들의 경험적 분포를 이용해 genomic inflation을 조절하는 Bayesian 방법으로, 450K와 EPIC 자료 모두에 적용해 λ를 1.0~1.1 수준으로 안정화시킬 수 있었습니다 [48].

(b) DMR (영역 기반 분석):
개별 CpG 단위의 EWAS 외에, 인접한 CpG들 간 **공동 변화 패턴**을 찾기 위해 DMR 분석을 수행합니다. DMR은 보통

EWAS 결과를 기반으로, 일정 거리 내 여러 CpG가 일관되게 차이가 나는 구간을 식별합니다. 사용되는 방법과 도구는 여러 가지가 있는데:

- **Bumphunter:** `bumphunter` 패키지는 미리 β값 스무딩을 거쳐, 클러스터된 CpG들에서 평균 차이가 임계치 이상인 "범프" 영역을 찾아냅니다 [49]. 입력으로 모델 행렬과 데이터를 주면 FWER 보정된 p값과 함께 DMR 리스트를 산출합니다. 다만 실행 시 permutation을 많이 돌려야 해서 시간이 걸릴 수 있습니다.
- **DMRcate: DMRcate** 패키지는 EWAS의 t값이나 확률값을 기반으로 공간상 가까운 CpG들을 합쳐 **Stouffer 통계** 등을 계산, DMR을 검출합니다. 비교적 사용이 간편하고 시각화 기능(`extractRanges`, `DMR.plot`)도 제공합니다.
- **ChAMP**의 DMR 기능: ChAMP 패키지의 `champ.DMR()` 함수는 bumphunter 엔진을 내부적으로 활용하여 DMR을 출력합니다. 또 ChAMP는 DMR 결과로부터 GSEA 준비까지 일련의 작업이 포함되어 있어 편리합니다 [50].
- **comb-p**: 파이썬 기반으로, EWAS p값의 **폭포 패턴**을 scanning하여 DMR을 찾는 도구 (메틸레이션 array에도 적용 가능). 하지만 R 통합이 부족하여 여기서는 주로 DMRcate나 bumphunter를 권장합니다.

DMR 결과는 **게놈 브라우저** 이미지로 시각화하면 효과적입니다. 예를 들어, **Figure**로 유의한 DMR 중 하나를 선정하여, 메틸레이션 β값 프로파일을 케이스 vs 컨트롤로 그려 줄 수 있습니다. DMRcate의 `DMR.plot()`을 이용하면 각 샘플의 β 분포와 유의구간을 그림으로 표시할 수 있습니다. 또한 DMR 목록을 **베노그램(Venn diagram)**으로 나타내, 예컨대 450K와 EPIC 각각 분석한 DMR의 교집합과 특이적 DMR 수를 비교하는 것도 가능합니다. EPICv2의 신규 영역에서만 발견되는 DMR이 있다면, 해당 DMR의 유전자나 기능 주석을 표로 정리하여 EPICv1 대비 **추가적 발견**을 강조할 수 있습니다.

(c) 결과 해석 및 검증:
발견된 DMP/DMR에 대해서는 **생물학적 의미**를 해석해야 합니다. 이를 위해 **Gene Ontology (GO) 및 경로 분석**을 수행할 수 있습니다. 예를 들어 ChAMP의 `champ.GSEA()`나 **missMethyl** 패키지의 `gometh()` 함수를 사용하면 EWAS 결과를 입력으로 과대대표된 GO term이나 경로를 알려줍니다. 흥미로운 결과는 **Table**로 정리하여, 예) "면역반응 관련 경로가 메틸레이션 변화에 유의하게 연결되었다" 등 기술합니다.

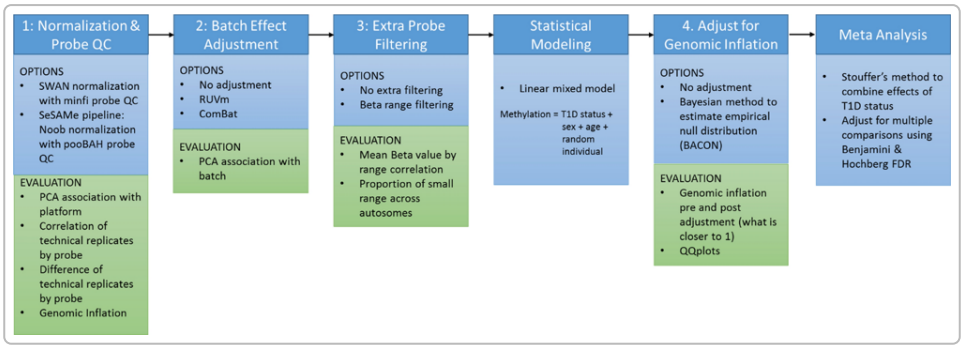
또한 **양성 컨트롤**을 다시 한 번 활용합니다. 만약 잘 알려진 EWAS 연관 CpG가 있다면 (예: 흡연-AHRR, 노화-ELOVL2 등), 본 결과에서 해당 CpG가 기대대로 나왔는지 검증합니다. 이것은 분석 파이프라인의 **유효성**을 보여주는 증거가 됩니다. 예컨대 "AHRR의 cg05575921이 본 연구 흡연자 vs 비흡연자 EWAS에서 가장 강력한 DMP로 검출되어, 이전 연구와 일관되게 확인되었다"라고 언급할 수 있습니다 [31].

마지막으로, DMP/DMR 중 일부는 **기능적 실험 데이터**로 교차검증할 수 있습니다. 예를 들어 특정 유전자의 프로모터 DMR이 발견됐다면, 해당 유전자의 발현 데이터나 HISTONE 마크 데이터와 연관성을 literature에서 찾아 언급하면 결과의 신뢰성을 높일 수 있습니다.

6. 시각화 및 품질 평가 (Visualization & Quality Assessment)

마지막 단계에서는 **종합적인 시각화**를 통해 전 과정의 품질과 결과를 요약합니다. 앞서 각 단계별로 제시한 그림 아이디어들을 논문 도식으로 구성하면 다음과 같습니다:

- **전체 파이프라인 흐름도 (Flowchart):** 분석 작업의 전반적인 순서를 도식화하여 제시합니다. 아래 그림은 이러한 파이프라인 개요의 한 예시입니다.



Illumina 450K/EPIC 데이터 통합 분석 파이프라인 요약. 각 단계별 선택 옵션과 평가 지표를 도식화하였다. 1단계에서 정규화 방법 (예: SWAN vs SeSAMe-Noob)과 프로브 QC 옵션들을 비교하고, PCA 및 기술복제 비교, genomic inflation 등을 평가한다. 2단계에서 배치 효과 교정 방법 (None, RUVm, ComBat)을 적용하고, PCA로 배치 연관성을 평가한다. 3단계에서 추가 프로브 필터링 여부를 결정하며, β값 분포 범위를 평가한다. 4단계에서 genomic inflation 보정(BACON 등) 적용 전후 λ값과 Q-Q플롯을 확인한다. 최종적으로 메타분석이나 통합 분석을 통해 결과를 산출한다.

- **QC/정규화 그림:** 샘플의 **베타값 분포밀도 곡선**(정규화 전후), **컨트롤 프로브 신호 그래프**, **Detection p-value 히트맵** 등을 Figure로 묶어 제시합니다. 예를 들어 Figure 1A에 정규화 전후 베타값 **density plot** (정규화로 Type I/II 분포가 잘 겹쳐졌는지 확인), 1B에 **PCA plot** (플랫폼 별 클러스터링이 정규화 후 완화되는지), 1C에 **기술복제 상관도** 등을 넣어 전처리 효과를 보여줄 수 있습니다.

- **배치 교정 그림:** Figure 2로 배치 효과 관련 플롯들을 제시합니다. 2A: **교정 전 PCA** (배치별 색상으로 클러스터 확인), 2B: **교정 후 PCA** (배치 구분 사라진 모습)으로 배치 효과 제거를 직관적으로 나타냅니다. 2C: **PVCA 바 차트**로 교정 전후 **분산 기여도 변화**를 그립니다 (예: 배치 요인 파란색 부분이 30%->5%로 감소) [24] . 2D: **kBET 결과**(배치 교정 전후 kBET acceptance rate 비교)나 **실루엣 지수 변화** 등을 도식화해도 좋습니다.

- **후성시계/생물학적 신호 그림:** Figure 3으로는 후성시계 및 주요 생물학적 효과를 보여줍니다. 3A: Horvath 등 **시계 예측치 vs 실제나이** 산점도 (색상으로 그룹 표시하여 그룹간 가속노화 여부 확인). 3B: **시계별 나이차이 박스플롯** (예: 환자군이 대조군보다 GrimAge 가속도가 높음을 *** 표시로 강조). 3C:** 양성 대조 CpG 메틸화도 **(예: AHRR cg05575921의 흡연자 vs 비흡연자 메틸화 차이를 박스플롯으로 나타내어, 기대대로 흡연자가 유의하게 낮음을 보여줌). 이러한 시각화는 독자에게 파이프라인이** 생물학적 중요한 신호를 잘 찾아냈다**는 메시지를 전달합니다.

- **EWAS/DMR 결과 그림:** Figure 4에서는 본격적인 결과를 요약합니다. 4A: **Manhattan plot**으로 EWAS p값 분포를 그려, 유의한 CpG들이 게놈 어느 영역에 몰려있는지 보여줍니다. 4B: **Q-Q plot**으로 관찰 vs 기대 p값을 그려, lambda값과 전반적인 통계적 양호도를 제시합니다 (교정 후 점들이 대각선에 가까움을 보여줌). 4C: **DMR 예시 브라우저 플롯** (앞서 언급한 유의 DMR 영역의 메틸화 패턴). 4D: **GO/pathway barplot** (상위 5개 정도의 유의 경로와 -log10(p) 값을 그래프로 표시). 이처럼 한 Figure에 다각도의 결과를 담아도 되고, 지면이 충분하다면 EWAS와 DMR을 별도 Figure로 나눌 수도 있습니다.

- **표 작성:** 표는 앞서 제시한 **배치 교정 전후 지표 표**(PVCA, kBET 등) 외에도, **상위 EWAS 히트**(Top DMP 10선의 목록), **주요 DMR 목록**, **시계별 상관 및 검정 표** 등을 포함합니다. 예를 들어 "표 2: 상위 10개 차등메틸화 CpG" 형태로, CpG ID, 이웃유전자, Chr 위치, Δβ, p값, FDR을 정리합니다. "표 3: 후성시계 상관관계" 형태로 Horvath, Hannum, 실제나이 간 피어슨 상관계수를 매트릭스로 제시할 수도 있습니다.

각 단계별 이러한 **시각화와 표를 통해**, 파이프라인의 효과를 독자에게 명확히 전달합니다. 특히 배치 효과 제거와 생물학적 신호 보존 간의 **균형**이 핵심 메시지이므로, 예시 그림에서 보였듯 배치 교정 전후 **나타나는 변화**와 **유지되는 신호**를 대비하여 보여주는 것이 중요합니다.

마지막으로, 결과 서술 시에는 **한계점 및 향후 개선**도 언급하면 좋습니다. 예를 들어 “본 파이프라인은 450K vs EPICv2 플랫폼 차이를 효과적으로 조정하였으나, 완전히 동일한 프로브 세트를 분석하는 것은 아니므로 극미한 플랫폼 차이는 완전히 배제할 수 없습니다”라든지, “추후 연구에서는 전체 게놈 시퀀싱 메틸레이션과의 비교를 통해 본 메틸레이션 어레이 결과의 포괄성을 평가해야 합니다” 등의 논평으로 글을 맺습니다.

以上の 구성으로, Prof. Kang의 요청에 부합하는 **유연하고 신뢰도 높은 DNA 메틸레이션 통합 분석 파이프라인**이 완성됩니다. 이 파이프라인은 다양한 플랫폼과 샘플을 포괄하면서도, 각 단계에서 검증과 최적화를 거쳐 **배치 효과를 제거하고 생물학적 신호를 보존**하는 것을 목표로 합니다. 이를 통해 도출된 메틸레이션 패턴과 후성나이 지표 등은 향후 임상 및 생물학 연구에 신뢰성 있게 활용될 수 있을 것입니다.

참고문헌: 각 단계별 언급한 문헌 및 사용 패키지에 대한 출처는 본문 중에 인용형태【】로 표시하였습니다. 주요 참고로 Vanderlinden 등 (2021)의 450K-EPIC 조화 파이프라인 연구 【9】【51】, Zhuang 등 (2025)의 EPICv2 vs v1 비교 연구 【17】, Peters 등 (2024)의 EPICv2 특성 분석 【5】, 그리고 DNAme 배치 통합 지표 관련 Nature Methods (2021) 논문 【25】 등을 들 수 있습니다.

---

1 3 9 10 11 13 14 15 20 33 34 45 46 47 48 51 An effective processing pipeline for harmonizing DNA methylation data from Illumina's 450K and EPIC platforms for epidemiological studies
https://d-nb.info/1246261677/34

2 4 5 6 7 8 30 Characterisation and reproducibility of the HumanMethylationEPIC v2.0 BeadChip for DNA methylation profiling
https://bmcgenomics.biomedcentral.com/counter/pdf/10.1186/s12864-024-10027-5.pdf?utm_source=consensus

12 Tutorial for EPIC v1 and EPIC v2 joint methylation data analysis in RnBeads
https://rnbeads.org/materials/data/tutorial/tutorial_EPICv1_v2/tutorial.html

16 17 19 37 38 39 40 Accounting for differences between Infinium MethylationEPIC v2 and v1 in DNA methylation-based tools - PubMed
https://pubmed.ncbi.nlm.nih.gov/39005299/

18 [PDF] Technical variability across the 450K, EPICv1, and EPICv2 DNA ...
https://pure.eur.nl/ws/portalfiles/portal/175440400/
Technical_variability_across_the_450K_EPICv1_and_EPICv2_DNA_methylation_arrays.pdf

21 25 26 Benchmarking atlas-level data integration in single-cell genomics | Nature Methods
https://www.nature.com/articles/s41592-021-01336-8?error=cookies_not_supported&code=250f481c-ff73-47b8-b504-629f39677817

22 23 24 pvcaBatchAssess function - RDocumentation
https://www.rdocumentation.org/packages/pvca/versions/1.12.0/topics/pvcaBatchAssess

27 Removing unwanted variation in a differential methylation analysis ...
https://pmc.ncbi.nlm.nih.gov/articles/PMC4652745/

28 ComBat-met: adjusting batch effects in DNA methylation data
https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqaf062/8136479

29 BEclear: Batch Effect Detection and Adjustment in DNA Methylation ...
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0159921

31 AHRR (cg05575921) hypomethylation marks smoking behaviour ...
https://pmc.ncbi.nlm.nih.gov/articles/PMC5520281/

32 Identification and evaluation of age-correlated DNA methylation ...
https://www.sciencedirect.com/science/article/abs/pii/S1872497316300461

35 DNA methylation GrimAge strongly predicts lifespan and healthspan
https://pmc.ncbi.nlm.nih.gov/articles/PMC6366976/

36 DNA methylation GrimAge version 2 - Aging-US
https://www.aging-us.com/article/204434/text

41 The wateRmelon User's Guide - Bioconductor
https://www.bioconductor.org/packages//release/bioc/vignettes/wateRmelon/inst/doc/wateRmelon.html

42 Potential reversal of biological age in women following an 8-week ...
https://www.aging-us.com/article/204602/text

43 Epigenetic GrimAge Test | Clock Foundation Projects
https://projects.clockfoundation.org/grimage-test

44 GrimAge Outperforms Other Epigenetic Clocks in the Prediction of ...
https://pmc.ncbi.nlm.nih.gov/articles/PMC8087266/

49 50 Methylation Array Data Analysis Tips
https://www.illumina.com/techniques/microarrays/methylation-arrays/methylation-array-data-analysis-tips.html