NGS library construction using fragmented/size selected DNA

There are several important considerations when preparing libraries from DNA samples.

the amount of starting material
whether the application is for resequencing
or de novo sequencing

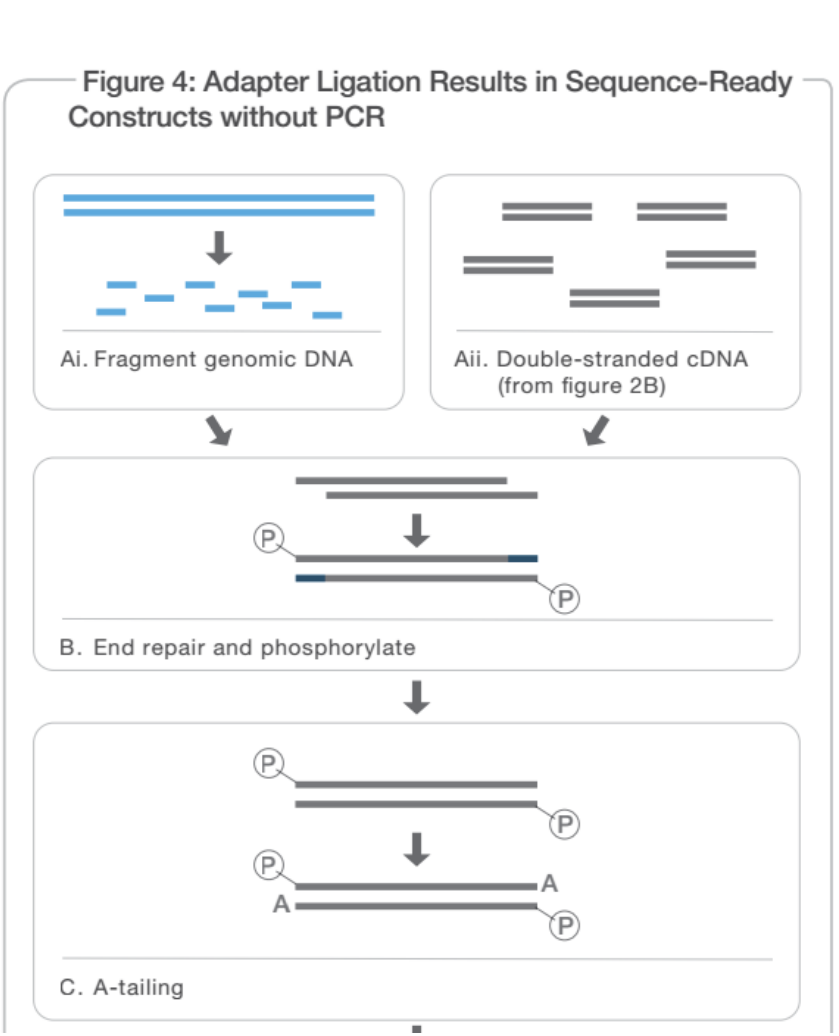
- Library preparations can be susceptible to bias resulting from genomes that contain unusually **high or low GC content**
- approaches have been developed to address these situations through careful selection of polymerases for PCR amplification, thermocycling, conditions and buffers

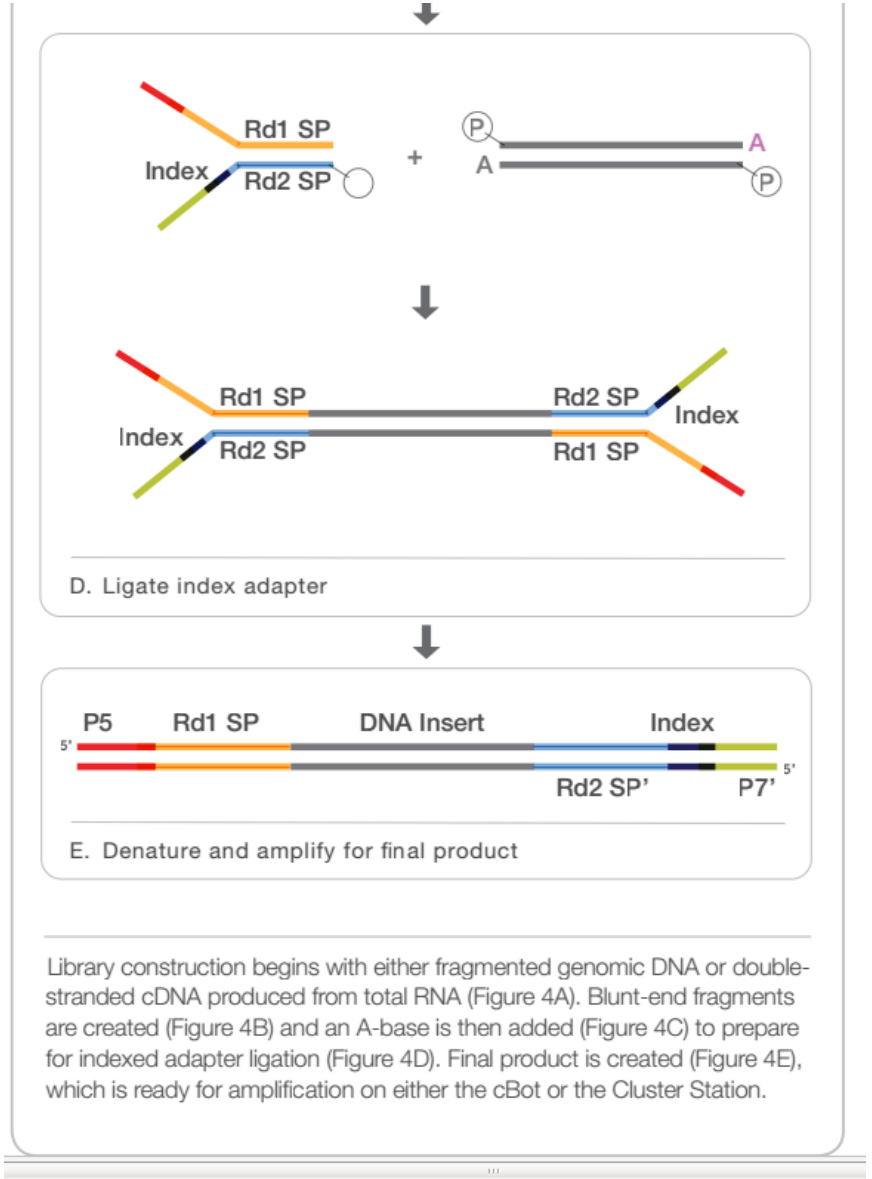
- Library preparation from DNA samples for sequencing
  - whole genomes,
  - targeted regions within genomes (for example exome sequencing),
  - ChIP-seq experiments,
  - or PCR amplicons follows the same general workflow.
- Ultimately, for any application, the goal is to make the libraries as complex as possible

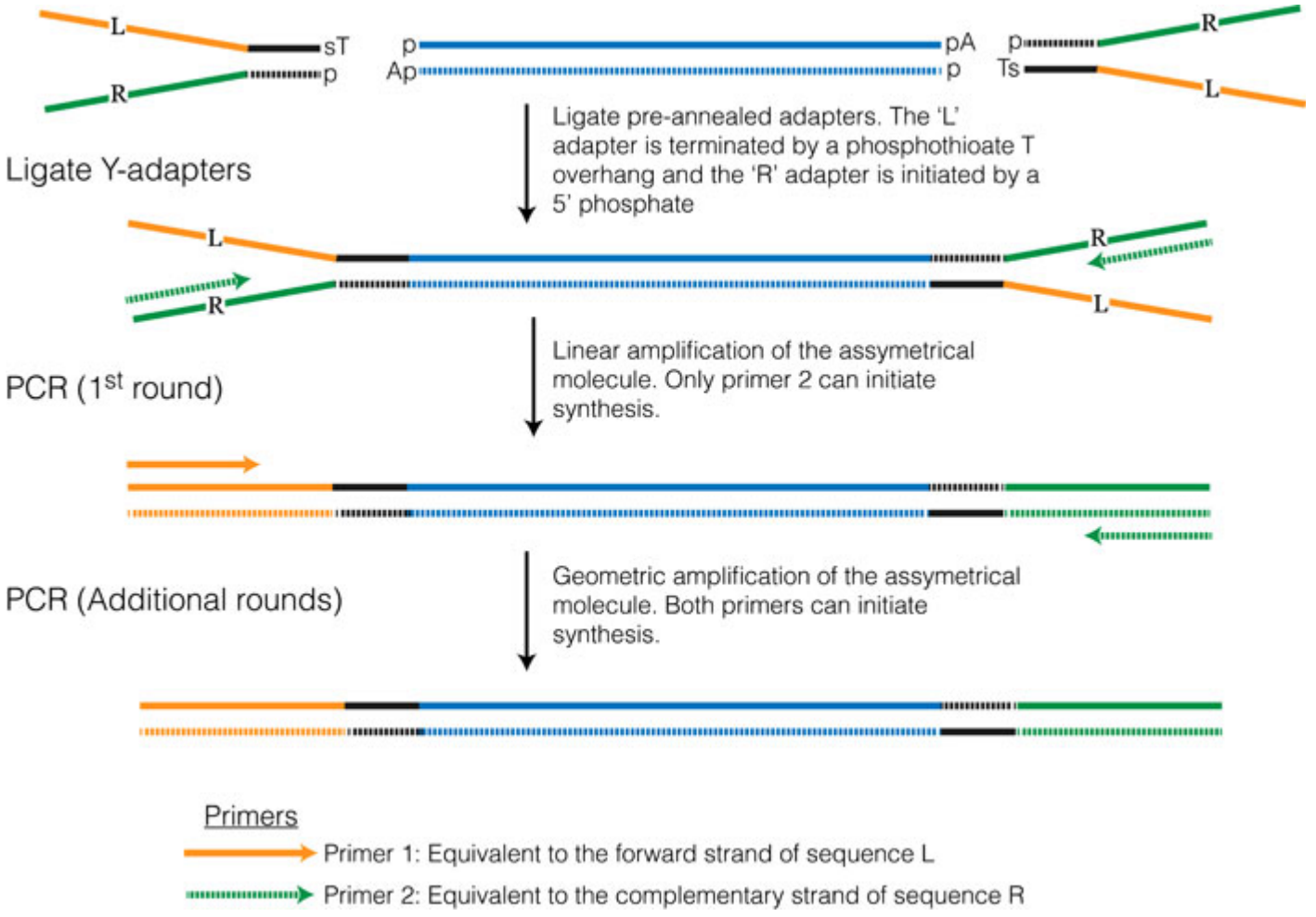

PARS-SEQ
SHAPE-SEQ
MDA
MALBAC
CHIRP-SEQ
RNA-SEQ
MAINE-SEQ
RRBS-SEQ
RIBO-SEQ
ICE
OS-SEQ
CLASH-SEQ
TIF-SEQ/PEAT
IN-SEQ
TAB-SEQ
CHIP-SEQ
5-C
SMMIP
OXBS-SEQ
RIP-SEQ
4-C
TC-SEQ
NET-SEQ
UMI
CAP-SEQ
FAIRE-SEQ
DUPLEX-SEQ
DNASE-SEQ
PAR-CLIP-SEQ
BS-SEQ
MEDIP-SEQ
GRO-SEQ
MERIP-SEQ
DIGITAL
CHIA-PET
ATAC-SEQ
CIP-TAP
ICLIP
HITS-CLIP
MBDCAP-SEQ
PARE-SEQ/GMUT
HI-C/3-C
TRAP-SEQ
RC-SEQ
FRAG-SEQ

- Numerous kits for making sequencing libraries from DNA are available commercially from a variety of vendors.
- Kits are available for making libraries from microgram down to picogram quantities of starting material.
- However, one should keep in mind the general principle that more starting material means less amplification and thus better library complexity.

TruSeq DNA Sample Preparation



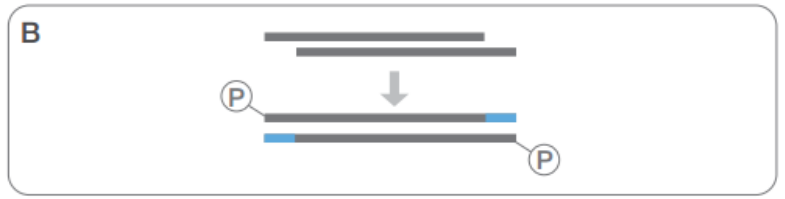



L
sT
R
p
p
pA
Ap
p
p
R
Ts
L
Ligate Y-adapters
Ligate pre-annealed adapters. The 'L' adapter is terminated by a phosphothioate T overhang and the 'R' adapter is initiated by a 5' phosphate
L
R
R
L
PCR (1st round)
Linear amplification of the assymetrical molecule. Only primer 2 can initiate synthesis.
PCR (Additional rounds)
Geometric amplification of the assymetrical molecule. Both primers can initiate synthesis.
Primers
Primer 1: Equivalent to the forward strand of sequence L
Primer 2: Equivalent to the complementary strand of sequence R

TruSeq® DNA PCR-Free Sample Preparation Kit

Figure 2: Adapter Ligation Results in Sequence-Ready Constructs without PCR

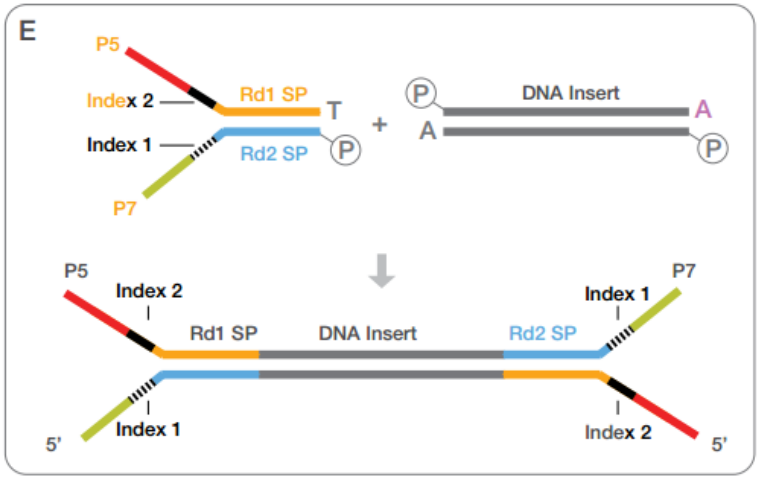
Library construction begins with genomic DNA that is subsequently fragmented



Blunt-end fragments are created.

Fragments are narrowly size selected with sample purification beads.

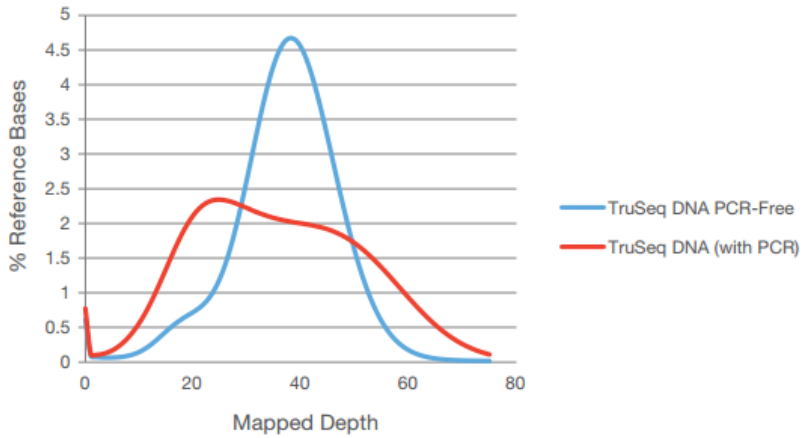
A-base is added.



Dual-index adapters are ligated to the fragments* and final product is ready for cluster generation.

*The TruSeq DNA PCR-Free LT indexing solution features a single-index adapter at this step.

Figure 4: Greater Coverage Uniformity

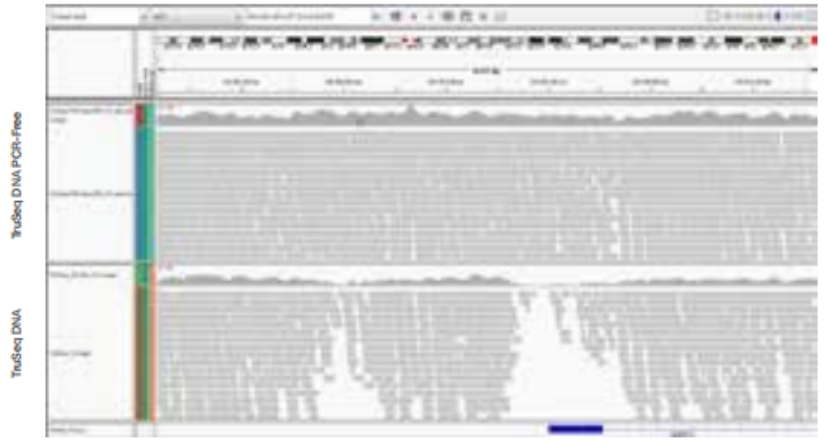


TruSeq DNA PCR-Free libraries provide greater coverage uniformity across the genome when compared to those generated using the TruSeq DNA protocol.

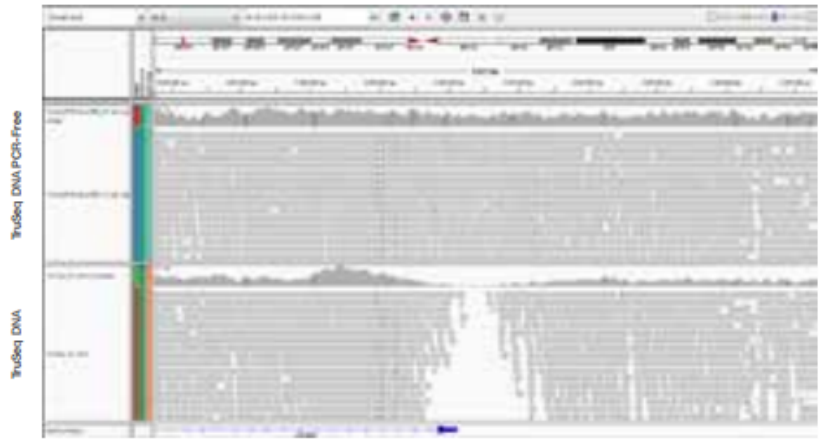
The PCR-Free kit also provides superior coverage of traditionally challenging genomic content, including GC-rich regions, promoters, and repetitive regions (Figure 5), allowing researchers to access more genomic information from each sequencing run (Figure 6).

Figure 6: PCR-Free Protocol Eliminates Coverage Gaps in GC-Rich Content

A



B

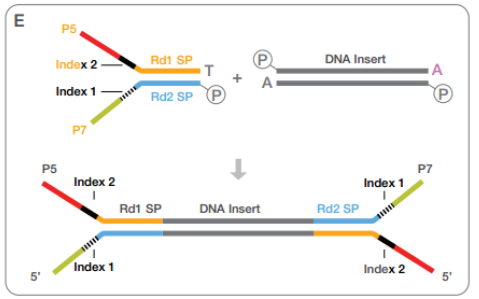


Increased coverage of TruSeq DNA PCR-Free libraries results in fewer coverage gaps, demonstrated here in the GC-rich coding regions of the *RNPEPL1* promoter (A) and the *CREBBP* promoter (B). PCR-Free sequence information is shown in the top panels of A and B, while sequence data generated using TruSeq DNA protocol (with PCR) are shown in the lower panels.

TruSeq® Nano DNA Sample

- The TruSeq Nano DNA protocol supports shearing by Covaris ultrasonication, requiring 100 ng of input DNA for an average insert size of 350 bp or 200 ng DNA for an average insert size of 550 bp.

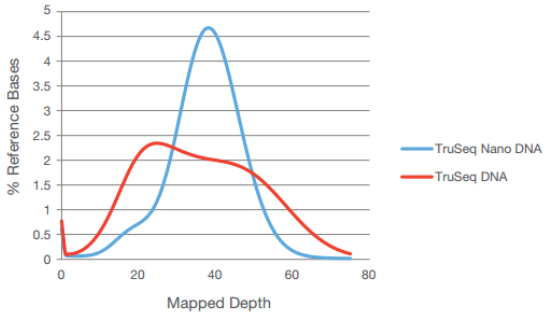
A-base is added.



Dual-index adapters are ligated to the fragments.

Ligated product is amplified and ready for cluster generation.

ıe TruSeq Nano DNA LT indexing solution features a single-index adapter
: Step E.

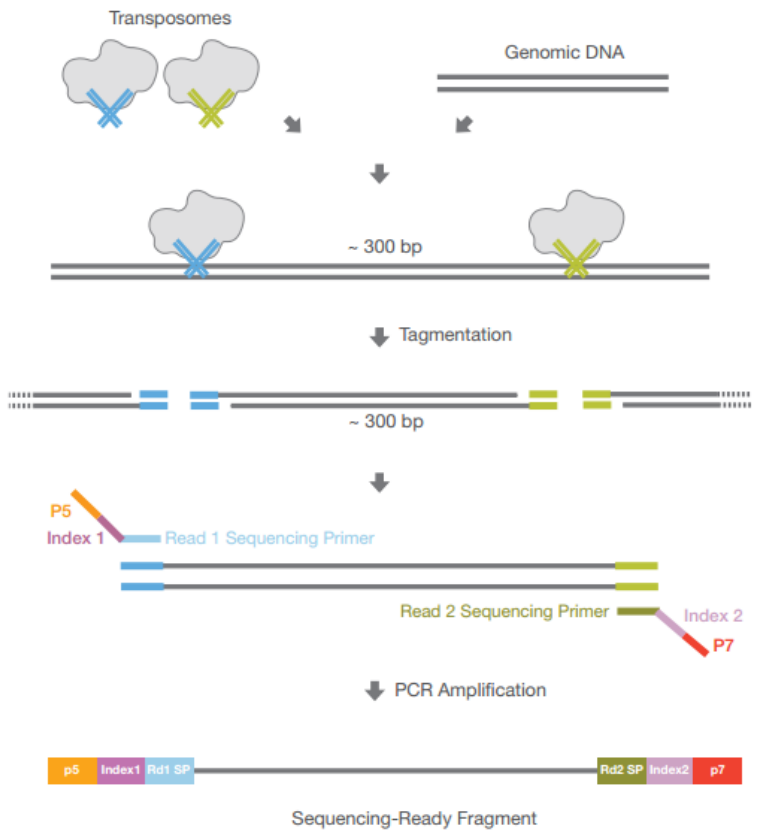


TruSeq Nano DNA libraries provide greater coverage uniformity across the genome when compared to those generated using the TruSeq DNA protocol.

Nextera®
DNA Sample Preparation Kits

- Easiest to Use
  - Prepare sequencing-ready samples in 1.5 hours with 15 minutes hands-on time
- Lowest DNA Input
  - Use just 50 ng DNA per sample, enabling use with samples in limited supply

Figure 2: Nextera Sample Preparation Biochemistry



Nextera chemistry simultaneously fragments and tags DNA in a single step. A simple PCR amplification then appends sequencing adapters and sample indexes to each fragment.

Table 2: Representative Nextera Applications

Examples of Nextera Applications
Large-genome resequencing
Small-genome resequencing
Amplicon resequencing
Clone or plasmid sequencing

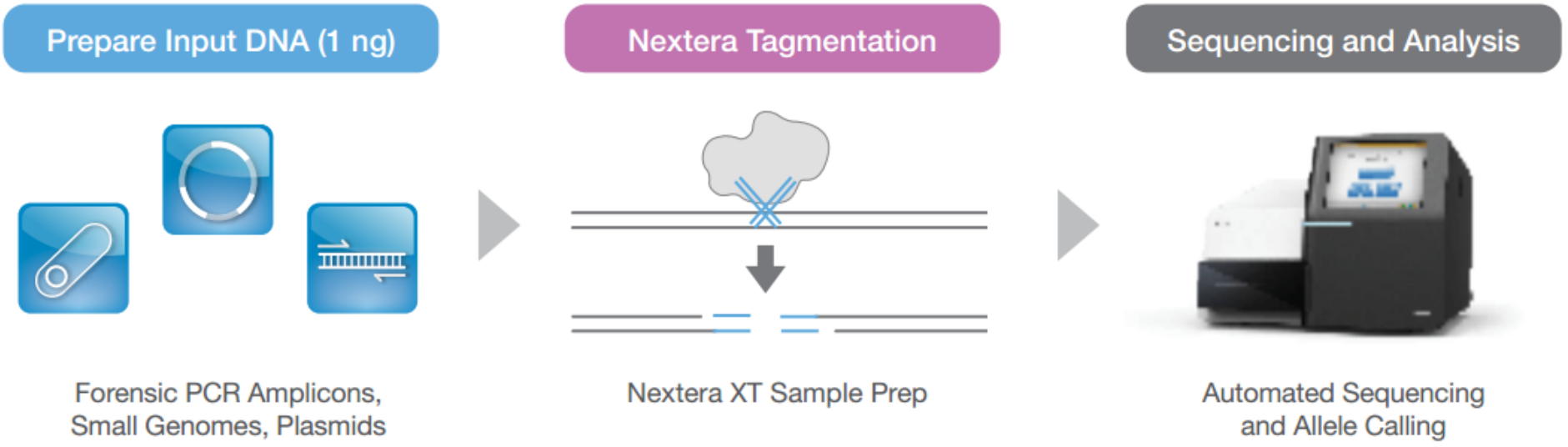
References

1. Ramirez MS, Adams MD, Bonomo RA, Centrón D, et al. (2011) Genomic analysis of *Acinetobacter baumannii* A118 by comparison of optical maps: Identification of structures related to its susceptibility phenotype. Antimicrob Agents Chemother, 55(4): 1520–6.
2. Adey A, Morrison HG, Asan, Xun X, Kitzman JO, et al. (2010) Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol 11: R119.
3. Bimber BN, Dudley DM, Lauck M, Becker EA, Chin EN, et al. (2010) Whole-genome characterization of human and simian immunodeficiency virus intrahost diversity by ultradeep pyrosequencing. J Virol 84: 12087–92.
4. Kitzman JO, Mackenzie AP, Adey A, Hiatt JB, Patwardhan RP, et al. (2010) Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nat Biotechnol 29: 59–63.
5. Linnarsson, S. (2010) Recent advances in DNA sequencing methods - General principles of sample preparation. Exp Cell Res 316: 1339–43.
6. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, et al. (2010) Diversity of human copy number variation and multicopy genes. Science 330: 641–646.
7. Voelkerding KV, Dames S, and JD Durtschi (2010) Next generation sequencing for clinical diagnostics-Principles and application to targeted resequencing for hypertrophic cardiomyopathy. J Mol Diagn 12: 539–551.

Accelerated Applications

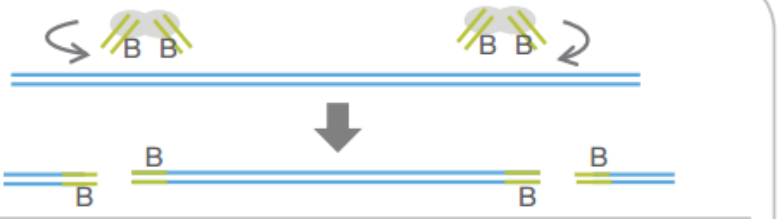
Nextera XT

Figure 1: Nextera XT Sample Preparation Workflow

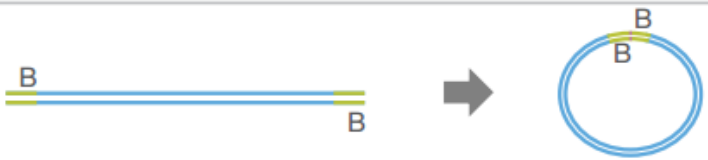


The combination of Nextera XT and rapid sequencing with the MiSeq System provides a complete DNA to data workflow in only 8 hours.

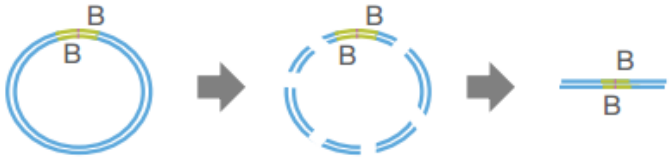
Figure 1: Nextera Mate Pair Workflow



Genomic DNA (blue) is tagmented with a Mate Pair Tagment Enzyme, which attaches a biotinylated junction adapter (green) to both ends of the tagmented molecule.



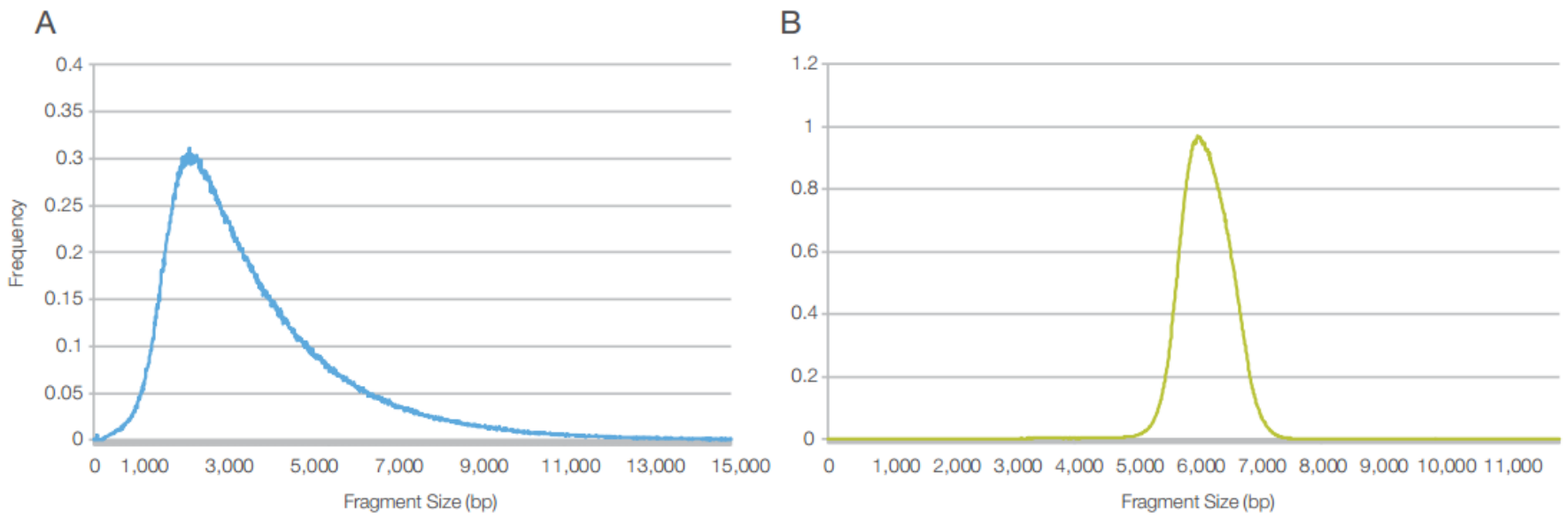
The tagmented DNA molecules are then circularized and the ends of the genomic fragment are linked by two copies of the biotin junction adapter.



Circularized molecules are then fragmented again, yielding smaller fragments. Sub-fragments containing the original junction are enriched via the biotin tag (B) in the junction adapter.

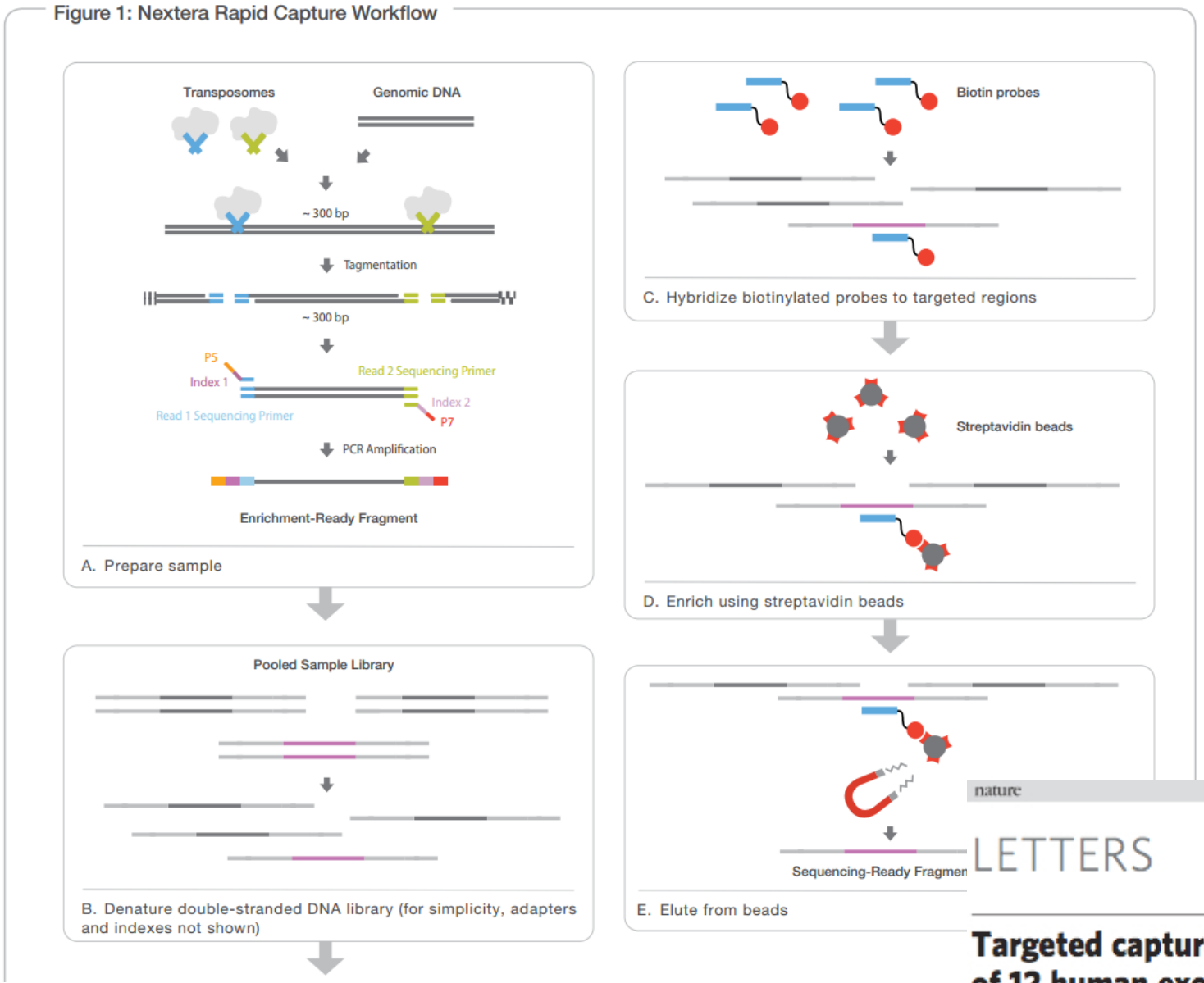
After End Repair and A-Tailing, TruSeq DNA adapters (gray and purple) are then added, enabling amplification and sequencing.

Figure 2: Fragment Size Distribution with Dual Protocols



Panel A shows the fragment size distribution of an *E. coli* mate pair library prepared using the Nextera Mate Pair gel-free protocol, resulting in a broad fragment size distribution. Panel B shows the narrow fragment size distribution of an *E. coli* mate pair library generated with the Nextera Mate Pair gel-plus protocol with automated size selection using the Pippin Prep platform.

Rapid Capture Exomes



nature Vol 461|10 September 2009 doi:10.1038/nature08250

LETTERS

Targeted capture and massively parallel sequencing of 12 human exomes

Sarah B. Ng[1], Emily H. Turner[1], Peggy D. Robertson[1], Steven D. Flygare[1], Abigail W. Bigham[2], Choli Lee[1], Tristan Shaffer[1], Michelle Wong[1], Arindam Bhattacharjee[4], Evan E. Eichler[1,3], Michael Bamshad[2], Deborah A. Nickerson[1] & Jay Shendure[1]

Target Capture Workflow

= target sequences
= off-target sequences
= adapter sequences
= biotin

Prepped library fragments

Add blocking oligos 1

xGen® Blocking Oligos

Blocked library fragments

Hybridize targets to capture probes 2

xGen® Lockdown® Probes

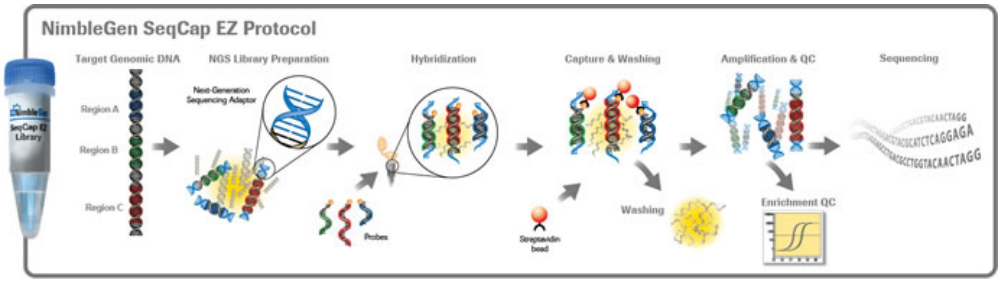
Hybridized library fragments

Incubate with magnetic streptavidin beads 3

Streptavidin-coated magnetic beads

Bead-bound target sequences

Isolate targets with magnet 4



IDT
INTEGRATED DNA TECHNOLOGIES

Focused investigation of key genes

With targeted resequencing, a subset of genes or regions of the genome are isolated and sequenced.

Targeted approaches using next-generation sequencing (NGS) allow researchers to focus time, expenses, and data analysis on specific areas of interest.

Such targeted analysis can include the exome (the protein-coding portion of the genome), specific genes of interest (custom content), targets within genes, or mitochondrial DNA.

Target enrichment:

Regions of interest are captured by hybridization to biotinylated probes and then isolated by magnetic pulldown. Target enrichment captures 20 kb–62 Mb regions, depending on the experimental design.

Ampliseq

Since dideoxy (Sanger) sequencing was developed over 30 years ago, amplicon sequencing has been a mainstay of genome analysis.

Now, with AmpliSeq™, it is possible to simultaneously amplify, sequence and genotype hundreds of genomic regions in a single project.

Ampliseq

By focusing next-generation DNA sequencing (NGS) technologies on specific targets, tens to hundreds of genetic markers can be quickly and cost effectively identified or genotyped in large populations.

video

Ion AmpliSeq™ Target Selection Overview

(RNA only) Reverse transcribe the RNA

↓

Amplify targets

↓

Partially digest primer sequences

↓

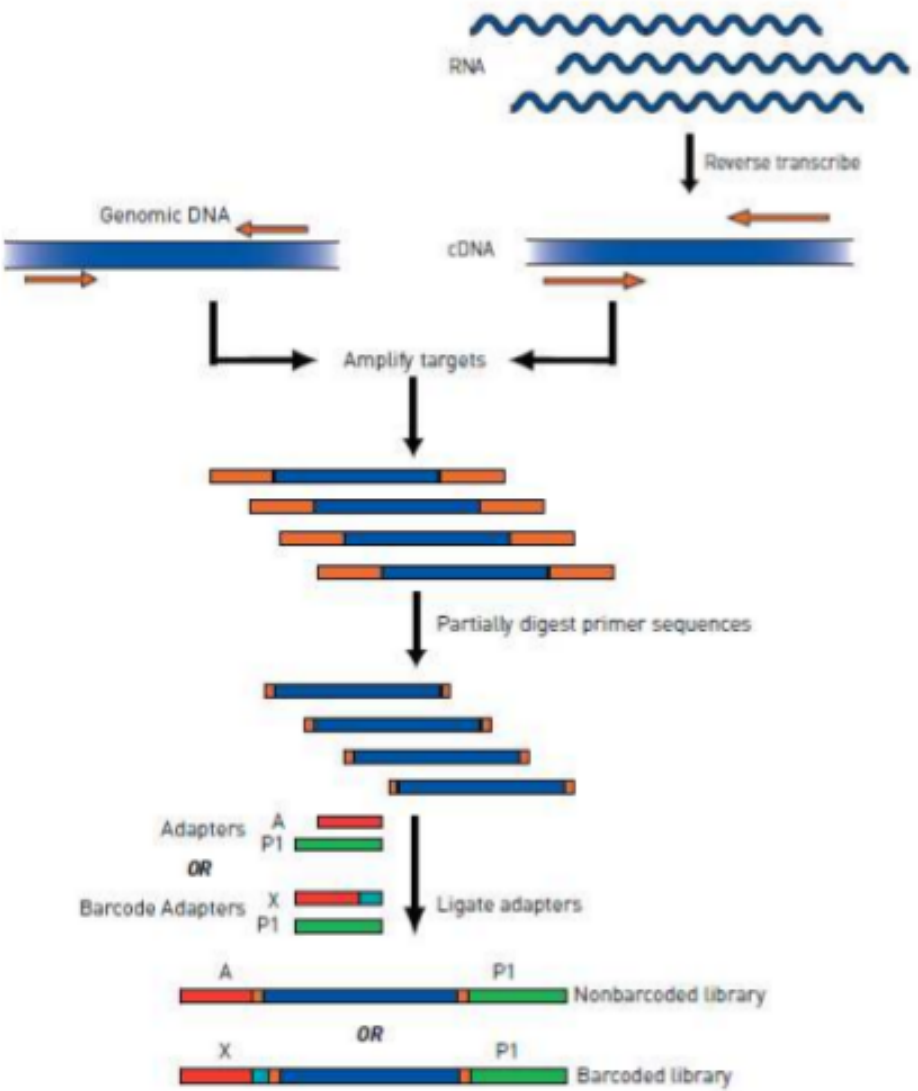
Ligate adapters to the amplicons and purify

↓

Option 1: Equalize the library (not recommended for libraries from RNA)	Option 2: Quantify the unamplified library by qPCR (recommended for libraries from RNA)	Option 3: Quantify the amplified library with the Qubit® 2.0 Fluorometer or Agilent® 2100 Bioanalyzer® instrument

↓

Combine libraries *(optional)* and proceed to template preparation.



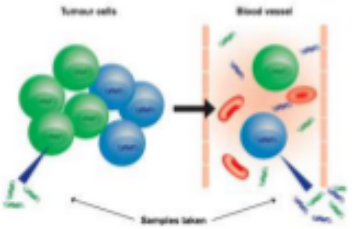
Key technology features:

- RNA pre-calculated designs
- DNA gene designs
  - "Standard" AmpliSeq designs
- DNA hotspot designs
  - 1-pool designs for SNP's

- Designs for "Any Genome"
  - AgBio applications and more

- Design customization
  - Sub-setting
  - Mix and match
  - Whitelist

- Support for Cell Free DNA (cfDNA) designs
  - 140bp amplicon sizes now available



Ion AmpliSeq™ Panels

For DNA analysis and RNA expression measurement

Ready-to-use **Human Identity Panel**	Ready-to-use **Human Ancestry Panel**	Ready-to-use **RNA Apoptosis Panel**	Ready-to-use **RNA Cancer Panel**
Ready-to-use **Cancer Hotspot Panel**	Ready-to-use **Comprehensive Cancer Panel**	Ready-to-use **Inherited Disease Panel**	Community **RNA Fusion Lung Cancer Research Panel**
Community ***BRCA* 1 & 2 Research Panel**	Community **Hearing Loss Research Panel**	Community ***CFTR* Research Panel**	Community ***TP53* Research Panel**
Community **AML Research Panel**	Community **Cardio Research Panel***	Community **Dementia Research Panel**	Community **Colon & Lung Cancer Research Panel**

Ion AmpliSeq™ Cancer Hotspot Panel v2

As little as 10ng input
50 genes
207 amplicons

The Ion AmpliSeq™ Cancer Panel targets 50 genes

ABL1	*EZH2*	*JAK3*	*PTEN*
AKT1	*FBXW7*	*IDH2*	*PTPN11*
ALK	*FGFR1*	*KDR*	*RB1*
APC	*FGFR2*	*KIT*	*RET*
ATM	*FGFR3*	*KRAS*	*SMAD4*
BRAF	*FLT3*	*MET*	*SMARCB1*
CDH1	*GNA11*	*MLH1*	*SMO*
CDKN2A	*GNAS*	*MPL*	*SRC*
CSF1R	*GNAQ*	*NOTCH1*	*STK11*
CTNNB1	*HNF1A*	*NPM1*	*TP53*
EGFR	*HRAS*	*NRAS*	*VHL*
ERBB2	*IDH1*	*PDGFRA*	
ERBB4	*JAK2*	*PIK3CA*	

Ion AmpliSeq™ Comprehensive Cancer Panel (CCP)

- Targets coding exons in 409 human oncogenes and tumor suppressor genes
- ~16,000 amplicons
- Detection of known COSMIC somatic mutations

Ion AmpliSeq™ Exome

- 293,903 primer pairs across 12 primer pools
  - >24,500-plex PCR!
  - ~2.4 M PCRs per plate of 8 exomes!!
- Total DNA input as low as 50ng
- Covers >97% of CCDS (Release 12)
  - >19,000 coding genes >198,000 coding exons (no UTRs, miRNAs, or ncRNAs)
  - ~85% of human disease-causing variants found in coding regions or splice junctions

- Amplicon size range 225-275 bp
  - Average insert size is ~202 bp

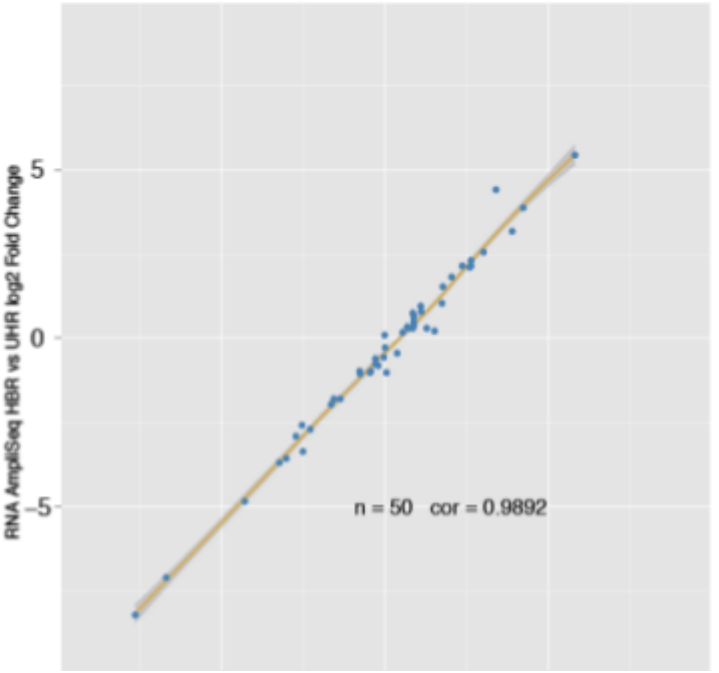
Ion AmpliSeq™ RNA Cancer Panel

50 genes, from 500 pg unfixed RNA, one tube

ABL1	EZH2	JAK3	PTEN	AKT1
FBXW7	IDH2	PTPN11	ALK	FGFR1
KDR	RB1	APC	FGFR2	KIT
RET	ATM	FGFR3	KRAS	SMAD4
BRAF	FLT3	MET	SMARCB1	CDH1
GNA11	MLH1	SMO	CDKN2A	GNAS
MPL	SRC	CSF1R	GNAQ	NOTCH1
STK11	CTNNB1	HNF1A	NPM1	TP53
EGFR	HRAS	NRAS	VHL	ERBB2
IDH1	PDGFRA	ERBB4	JAK2	PIK3CA

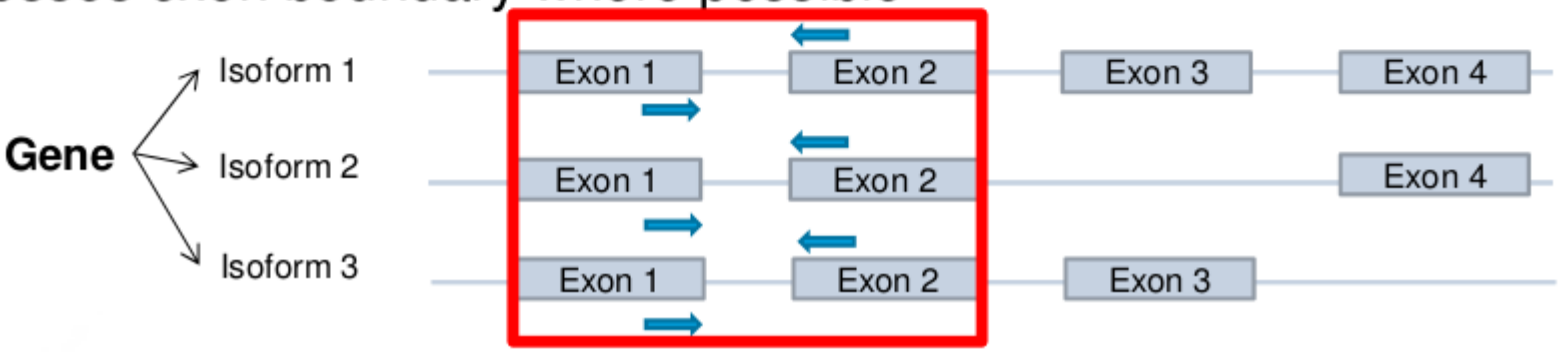
Targeted quantitative expression

Complement to Cancer Hotspot Panel



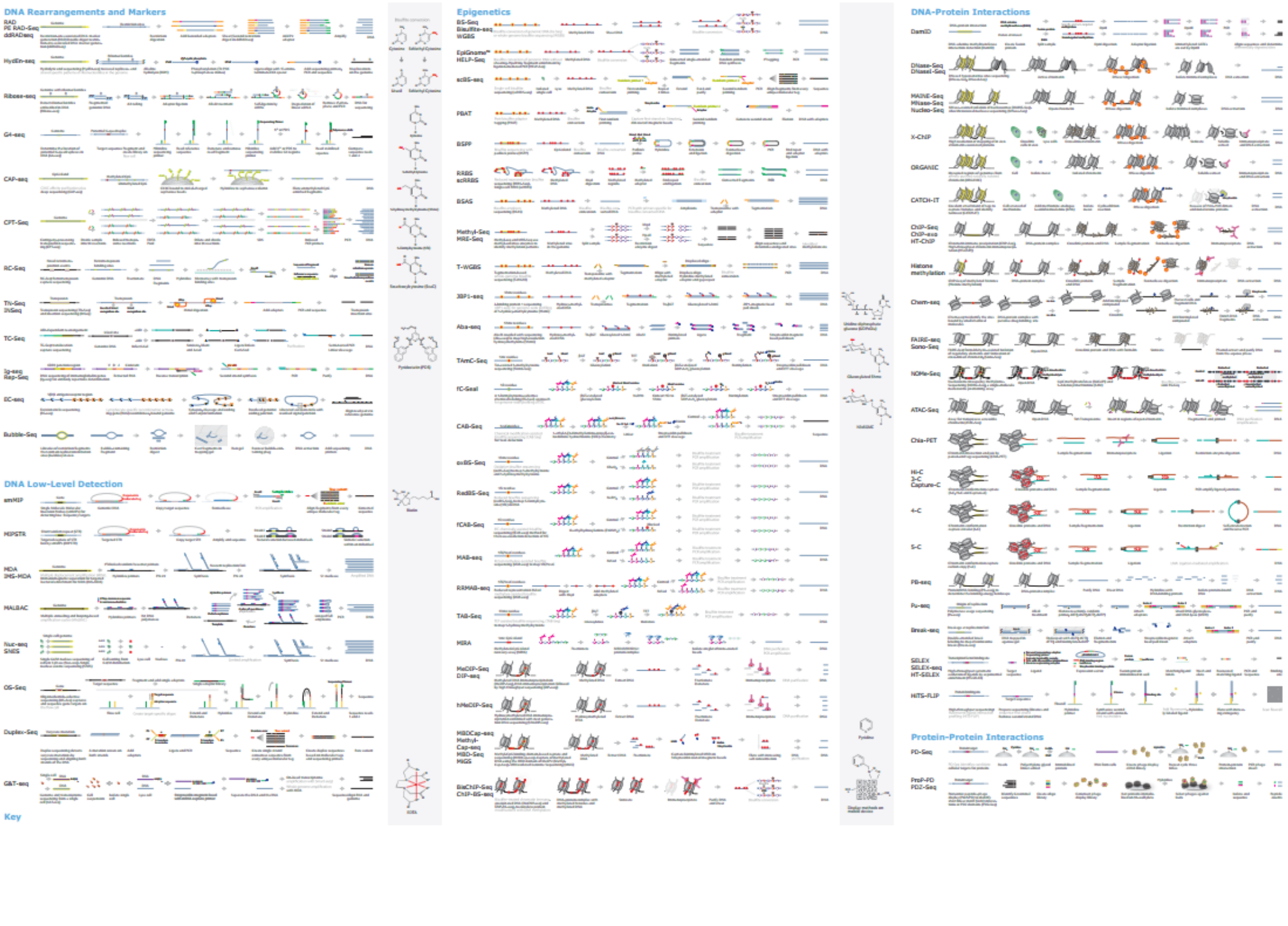
Ion AmpliSeq™ Transcriptome – Overview

- **Comprehensive** coverage of RefSeq - 20,802 genes targeted
  - Single primer pool – simple workflow
  - Reports at **gene-level**
- **FFPE-compatible** as little as 10 ng RNA
  - Total RNA input from FFPE and other sources – **no selection or enrichment**
- Assay design
  - One amplicon per gene
    - ~150 bp amplicon size with ~110 bp insert size
  - Crosses exon boundary where possible



Benefits of Target Enrichment vs. Amplicon Sequencing

Target Enrichment	Amplicon Sequencing
Larger gene content, typically > 50 genes	Smaller gene content, typically < 50 genes
More comprehensive profiling for all variant types	Ideal for analyzing single nucleotide variants and insertions/deletions (indels)
More comprehensive method, but with longer hands-on time and turnaround time	More affordable, easier workflow


DNA Rearrangements and Markers
RAD
PE RAD-Seq
ddRADseq
HydEn-seq
Ribose-seq
G4-seq
CAP-seq
CPT-Seq
RC-Seq
TN-Seq
INSeq
TC-Seq
Ig-seq
Rep-Seq
EC-seq
Bubble-Seq
DNA Low-Level Detection
smMIP
MIPSTR
MDA
IMS-MDA
MALBAC
Nuc-seq
SNES
OS-Seq
Duplex-Seq
GAT-seq
Key
Epigenetics
BS-Seq
Bisulfite-seq
WGBS
EpiGnome™
HELP-Seq
scBS-seq
PBAT
BSPP
RRBS
scRRBS
BSAS
Methyl-Seq
MRE-Seq
T-WGBS
3BP1-seq
Aba-seq
TAmC-Seq
fC-Seal
CAB-Seq
oxBS-Seq
RedBS-Seq
fCAB-Seq
MAB-seq
RRMAB-seq
TAB-Seq
MIRA
MeCEP-Seq
DIP-seq
hMeDIP-Seq
MBDCap-seq
Methyl-Cap-seq
MBD-Seq
MIGS
BisChIP-Seq
ChIP-BS-seq
DNA-Protein Interactions
DamID
DNase-Seq
DNaseI-Seq
MAINE-Seq
MNase-Seq
Nucleo-Seq
X-ChIP
ORGANIC
CATCH-IT
ChIP-Seq
ChIP-exo
HT-ChIP
Histone methylation
Chem-seq
FAIRE-seq
Sono-Seq
NOMe-Seq
ATAC-Seq
Chia-PET
Hi-C
3-C
Capture-C
4-C
5-C
PB-seq
Pu-seq
Break-seq
SELEX
SELEX-seq
HT-SELEX
HITS-FLIP
Protein-Protein Interactions
PD-Seq
PreP-PD
PDZ-Seq

LOW level DNA detection

Single-cell genomics can be used to identify and study circulating tumor cells, cell-free DNA, microbes, uncultured microbes, for preimplantation diagnosis, and to help us better understand tissue-specific cellular differentiation.

DNA replication during cell division is not perfect; as a result, progressive generations of cells accumulate unique somatic mutations.

Consequently, each cell in our body has a unique genomic signature, which allows the reconstruction of cell lineage trees with very high precision.

These cell lineage trees can predict the existence of small populations of stem cells. This information is important for fields as diverse as cancer development preimplantation, and genetic diagnosis.

Reviews

Blainey P. C. (2013) The future is now: single-cell genomics of bacteria and archaea. FEMS Microbiol Rev 37: 407-427

Lovett M. (2013) The applications of single-cell genomics. Hum Mol Genet 22: R22-26

Shapiro E., Biezuner T. and Linnarsson S. (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 14: 618-630

Low-Level DNA Detection

- MULTIPLE DISPLACEMENT AMPLIFICATION (MDA)

Multiple displacement amplification (MDA) is a method commonly used for sequencing microbial genomes due to its ability to amplify templates larger than 0.5 Mbp, but it can also be used to study genomes of other sizes.

- In this method, 3’-blocked random hexamer primers are hybridized to the template, followed by synthesis with Phi 29 polymerase.
- Phi 29 performs strand-displacement DNA synthesis, allowing for efficient and rapid DNA amplification.


3′ blocked random hexamer primers
Nascent replication fork
Hybridize primers
Phi 29
Synthesis
Phi 29
Synthesis
S1 nuclease

DNA-PROTEIN INTERACTIONS

Chromatin remodeling is a dynamic process driven by factors that change DNA-protein interactions. These epigenetic factors can involve protein modifications, such as histone methylation, acetylation, phosphorylation, and ubiquitination.

Histone modifications determine gene activation by recruiting regulatory factors and maintaining an open or closed chromatin state. Epigenetic factors play roles in tissue development, embryogenesis, cell fate, immune response, and diseases such as cancer. Bacterial pathogens can elicit transcriptional repression of immune genes by chromatin remodeling.

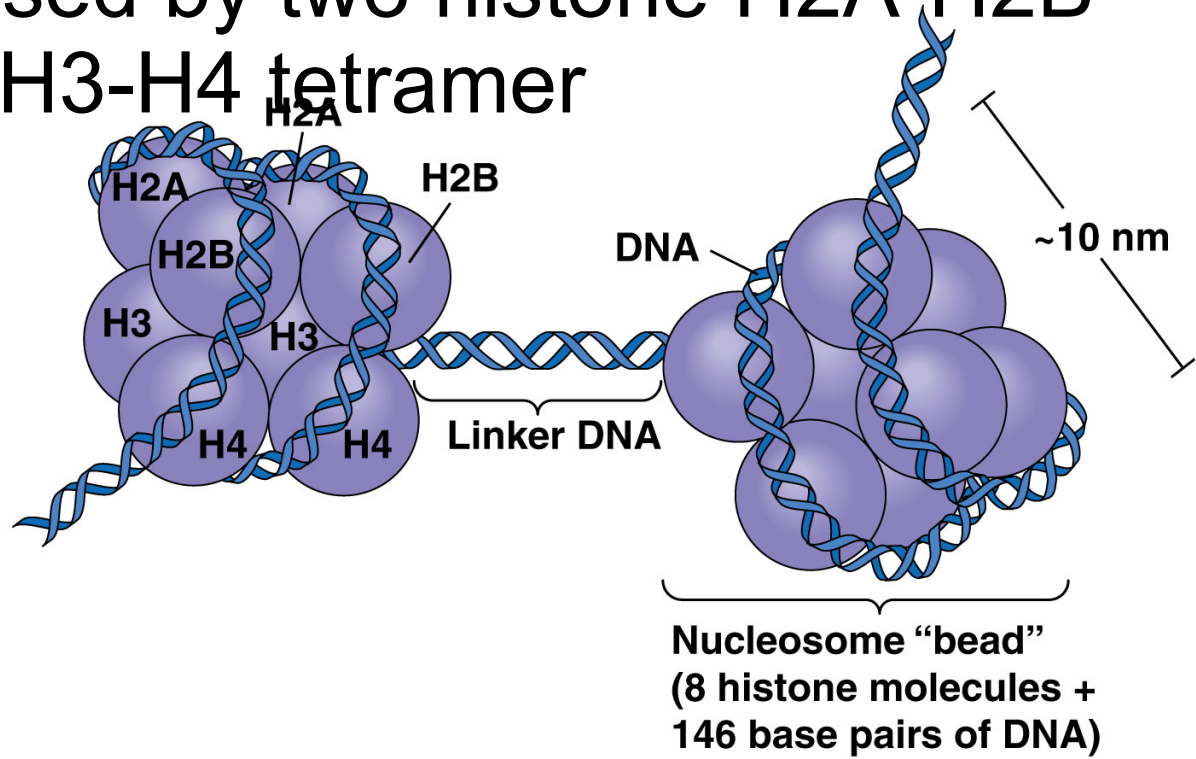
Cigarette smoking disrupts DNA-protein interactions leading to the development of cancers or pulmonary diseases.

DNASE I HYPERSENSITIVE SITES SEQUENCING (DNASE-SEQ)

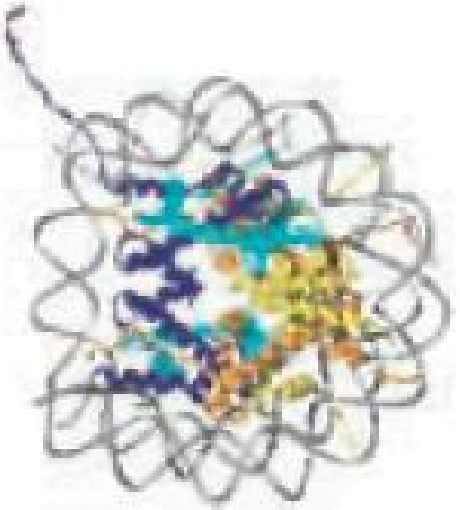
Sequences bound by regulatory proteins are protected from DNase I digestion. Deep sequencing provides accurate representation of the location of regulatory proteins in genome.

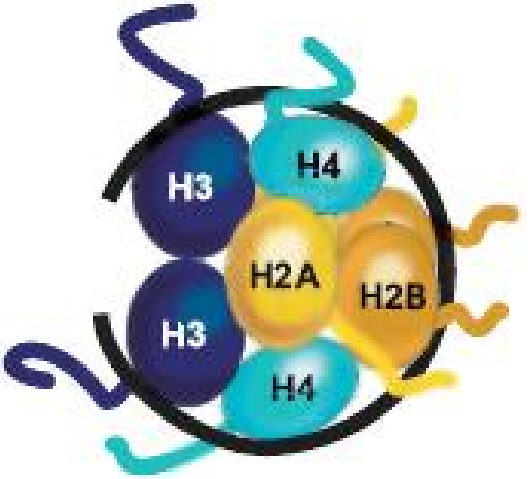
Chip-seq

In eukaryotic cells, the genome is highly organized within the nucleus in a complex compact structure known as chromatin. The basic unit of chromatin is the nucleosome, which consists of 146 bp of DNA wrapped around the four histone proteins arranged as an octamer composed by two histone H2A-H2B dimers and a histone H3-H4 tetramer



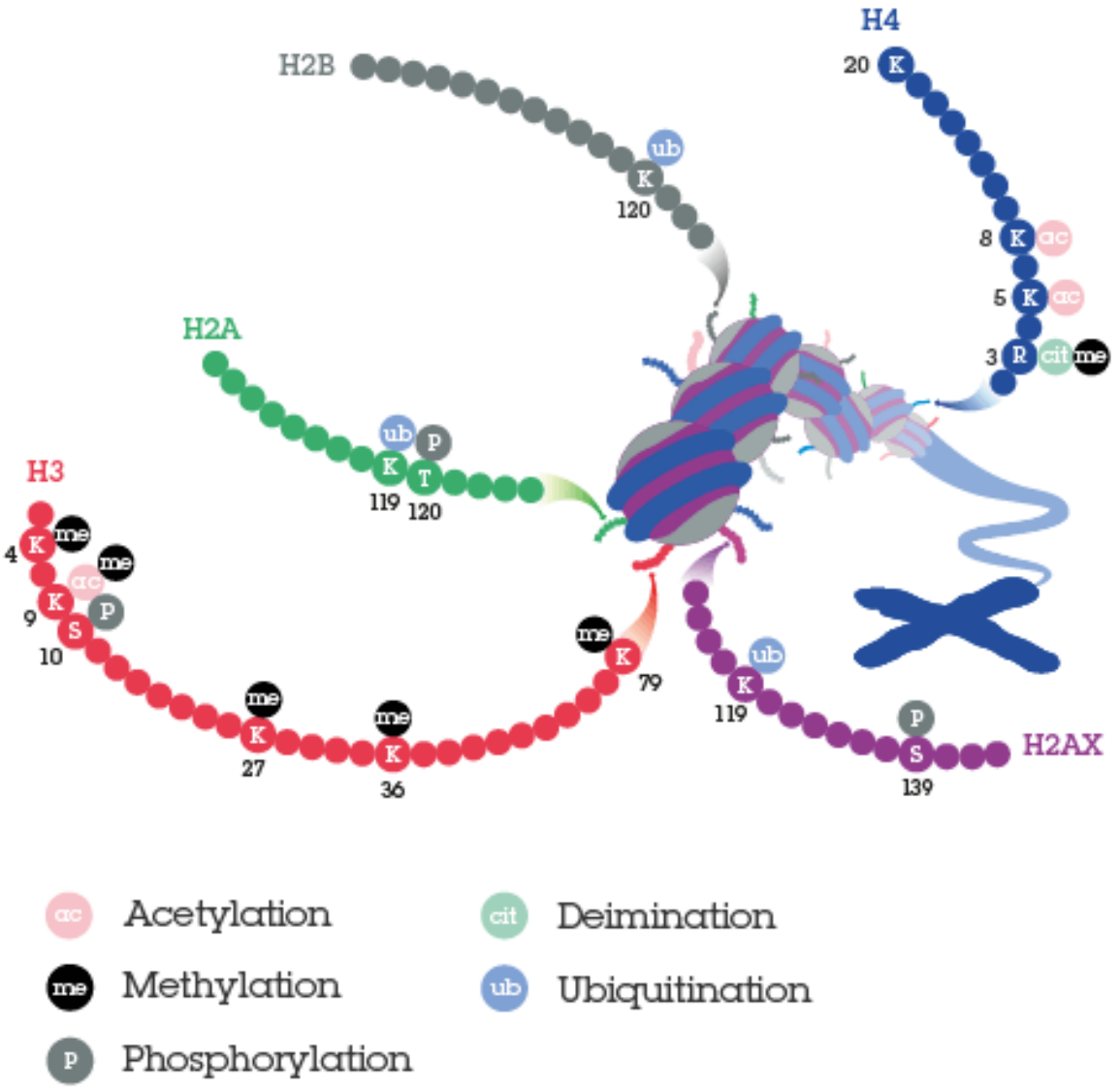
Each histone protein contain the so called histone fold structural motif, anked by unstructured N- and C-terminal tails, ranging from 15 (H2A) to 35 (H3) amino acids, that protrude from the nucleosomal core





Histone modifications

Acetylation and methylation of specic lysine or arginine residues in histones H3 and H4 are reversible and have been associated to gene transcription regulation



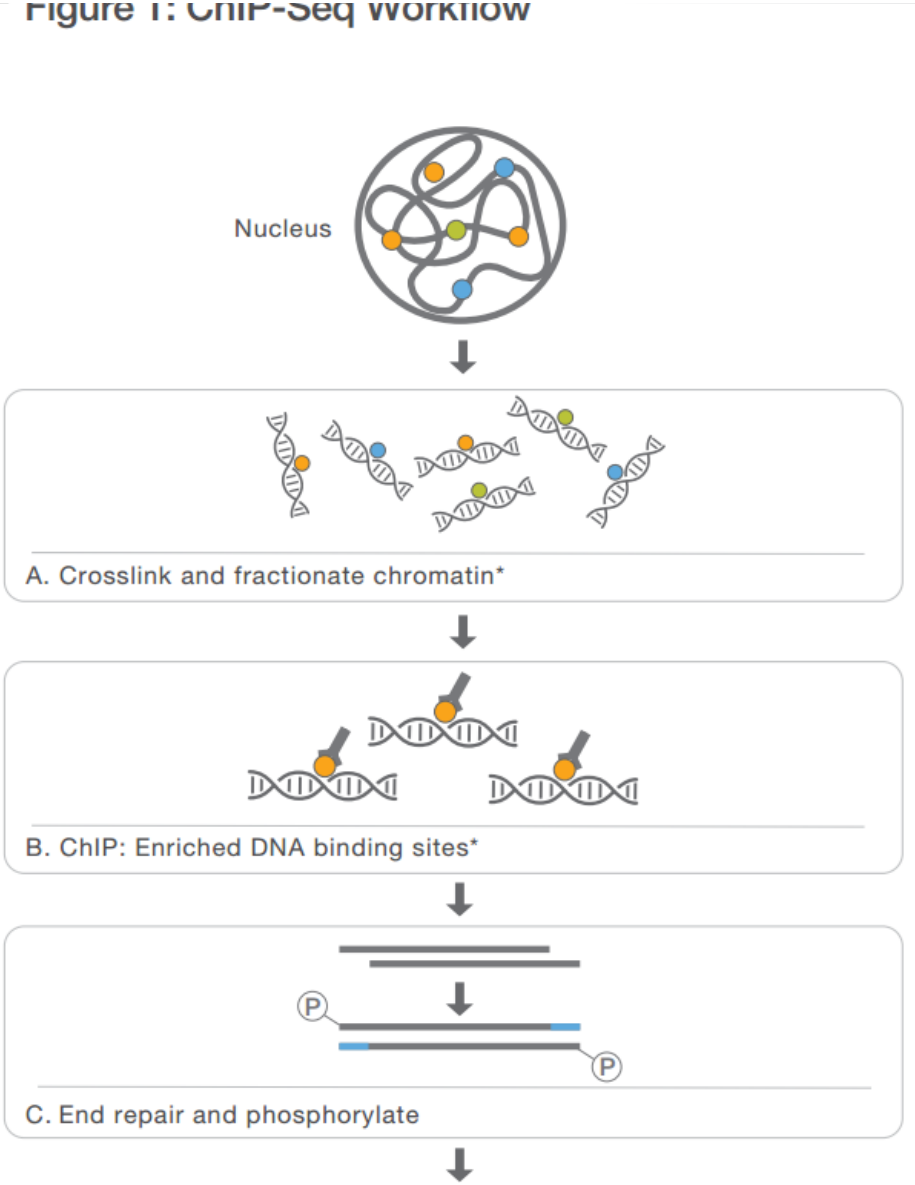
H3K4, H3K36, and H3K79 methylations are generally linked to active gene expression, whereas H3K9, H3K27, and H4K20 di- and tri-methylations have been associated with gene silencing.

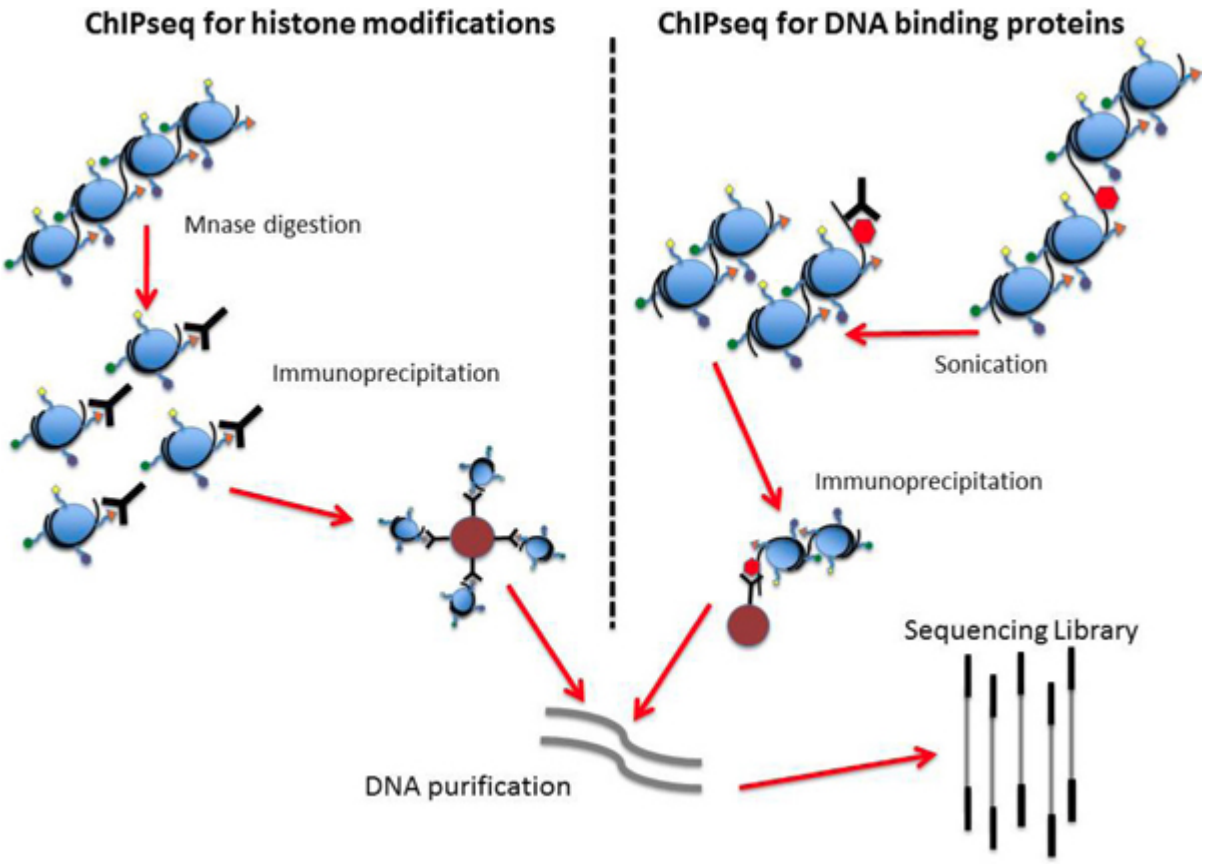
H3K4me2 and H3K4me3 have been both found predominantly on active loci, although H3K4me3 is associated with active genes, H3K4me2 can be present also in inactive genes

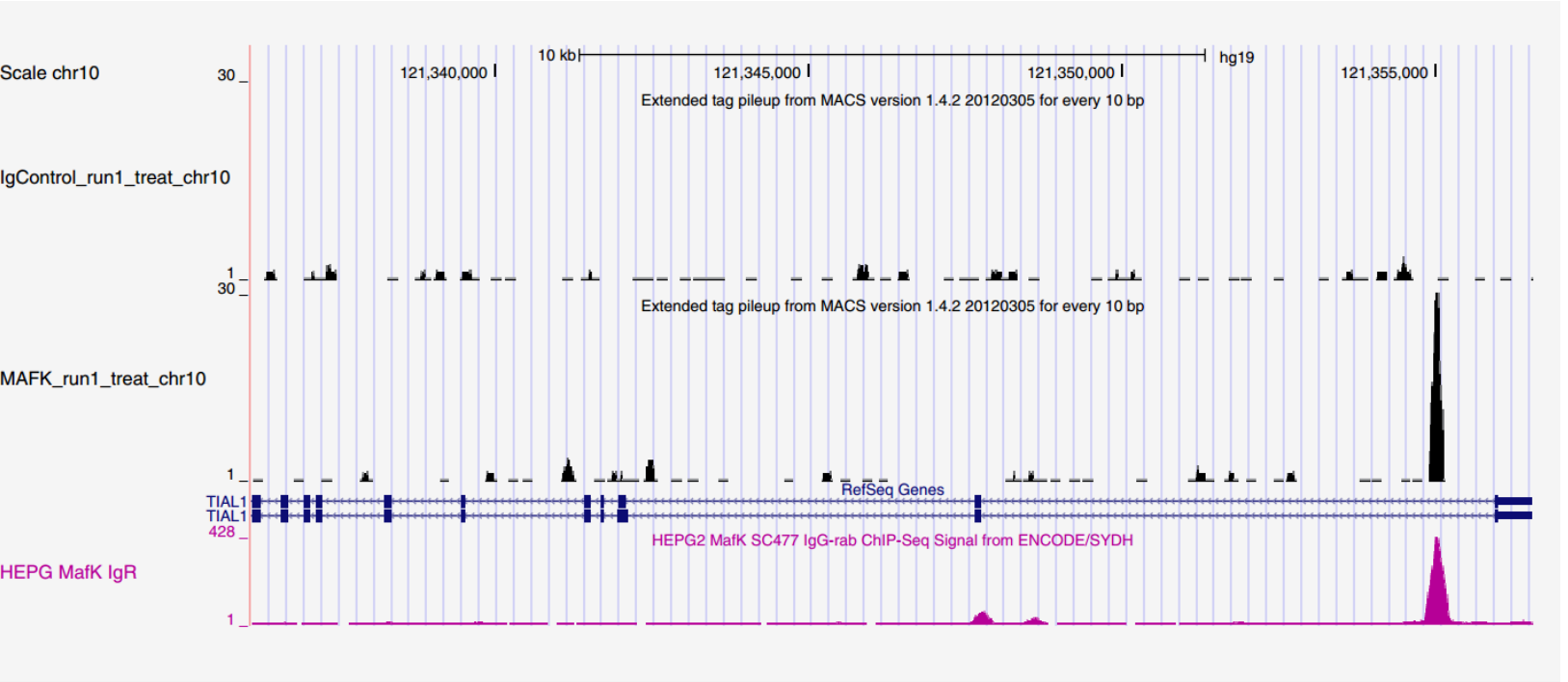
TruSeq® ChIP Sample Preparation

Determining how protein–DNA interactions regulate gene expression is essential for fully understanding many biological processes and disease states. This epigenetic information is complementary to DNA sequencing, genotyping, gene expression, and other forms of genomic analysis. Chromatin immunoprecipitation sequencing (ChIP-Seq) leverages next-generation sequencing (NGS) to quickly and efficiently determine the distribution and abundance of DNA-bound protein targets of interest across the genome. ChIP-Seq has become one of the most widely applied NGS-based applications, enabling researchers to reliably identify binding sites of a broad range of targets across the entire genome with high resolution and without constraints.

Figure 1: ChIP-Seq Workflow

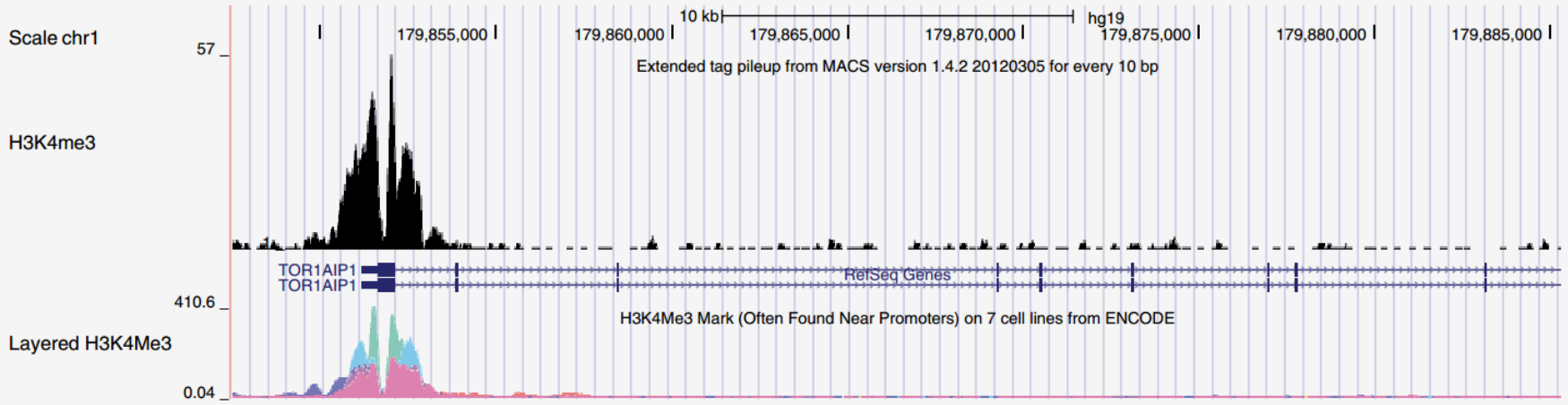



ChIPseq for histone modifications
ChIPseq for DNA binding proteins
Mnase digestion
Immunoprecipitation
Sonication
Immunoprecipitation
Sequencing Library
DNA purification

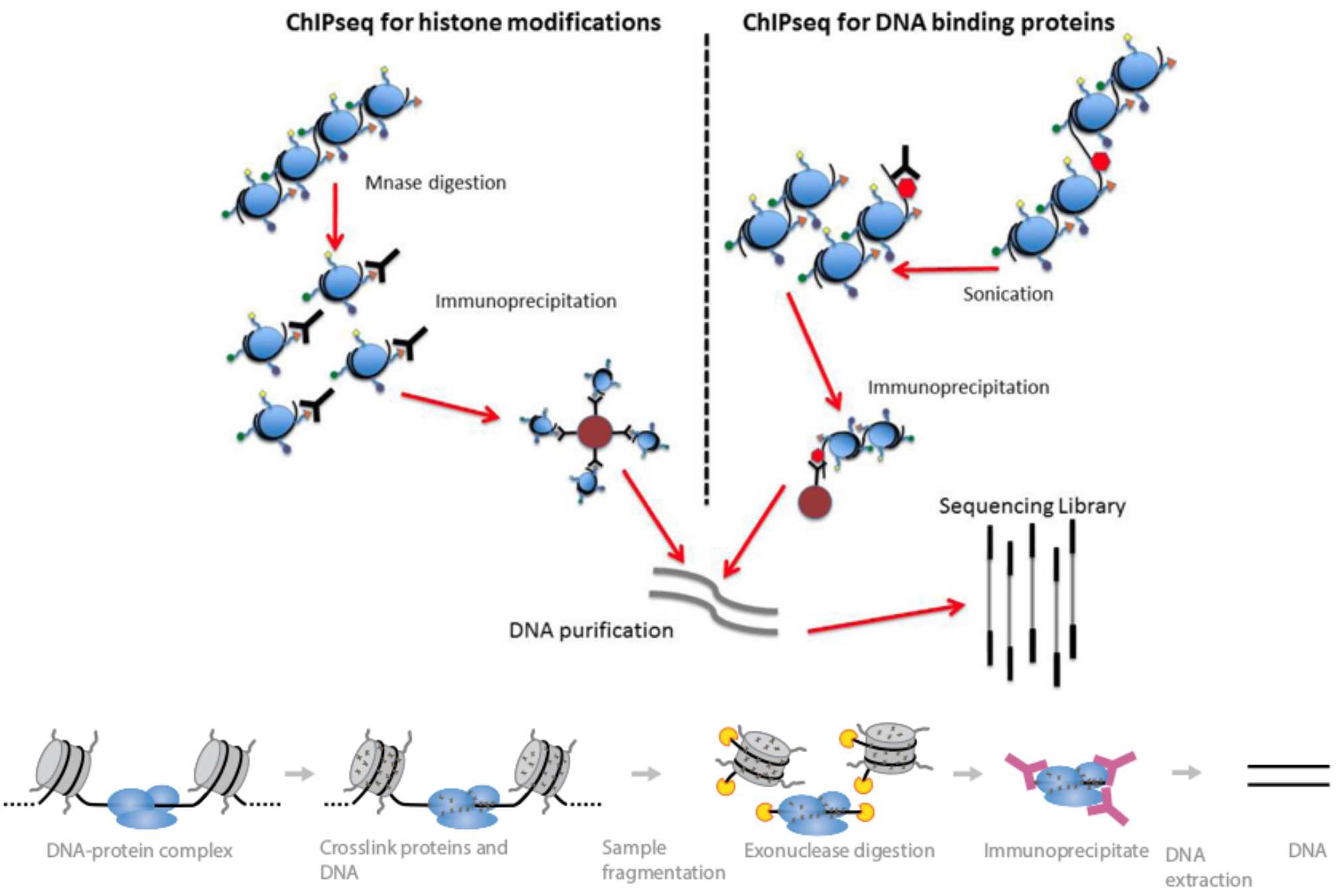


TruSeq ChIP Sample Preparation Kits enable the generation of libraries across a broad range of study designs. Above is peak data for a negative Ig control, the transcription factor target MafK, and a reference peak for MafK from the ENCODE database.

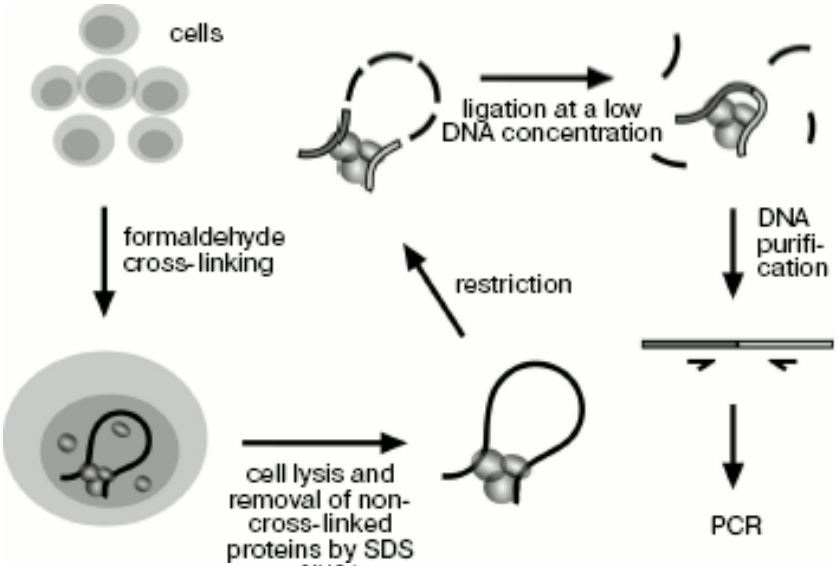
https://en.wikipedia.org/wiki/MAFK



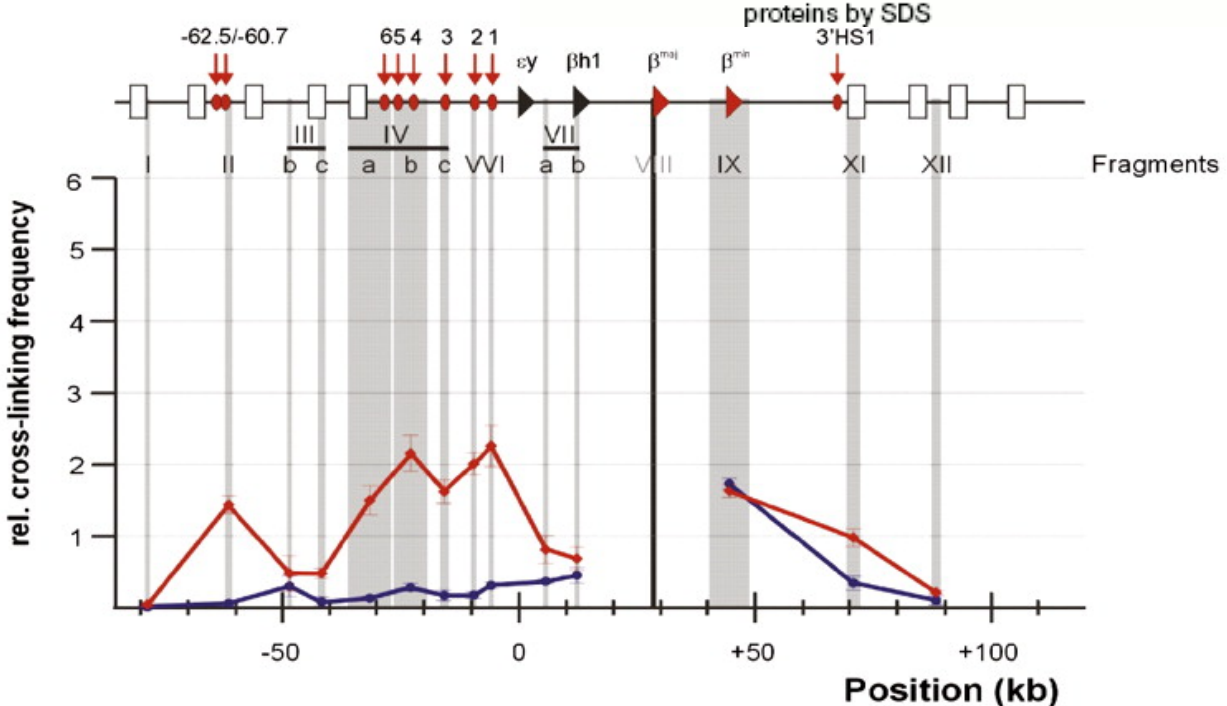
The peak results for the H3K4me3 target compare favorably with the ENCODE annotation data for this well characterized target, with a representative peak for the histone mark target H3K4me3 and a corresponding ENCODE reference peak.


ChIPseq for histone modifications
ChIPseq for DNA binding proteins
Mnase digestion
Immunoprecipitation
Sonication
Immunoprecipitation
Sequencing Library
DNA purification
DNA-protein complex
Crosslink proteins and DNA
Sample fragmentation
Exonuclease digestion
Immunoprecipitate
DNA extraction
DNA

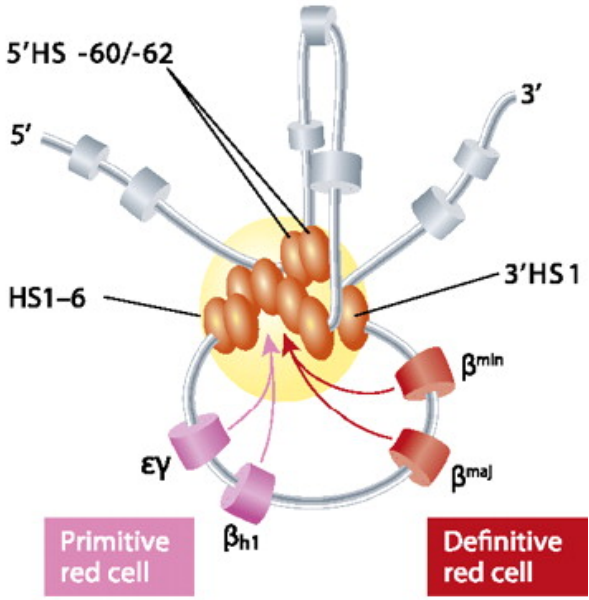
CHROMATIN CONFORMATION CAPTURE (HI-C/3C-SEQ)



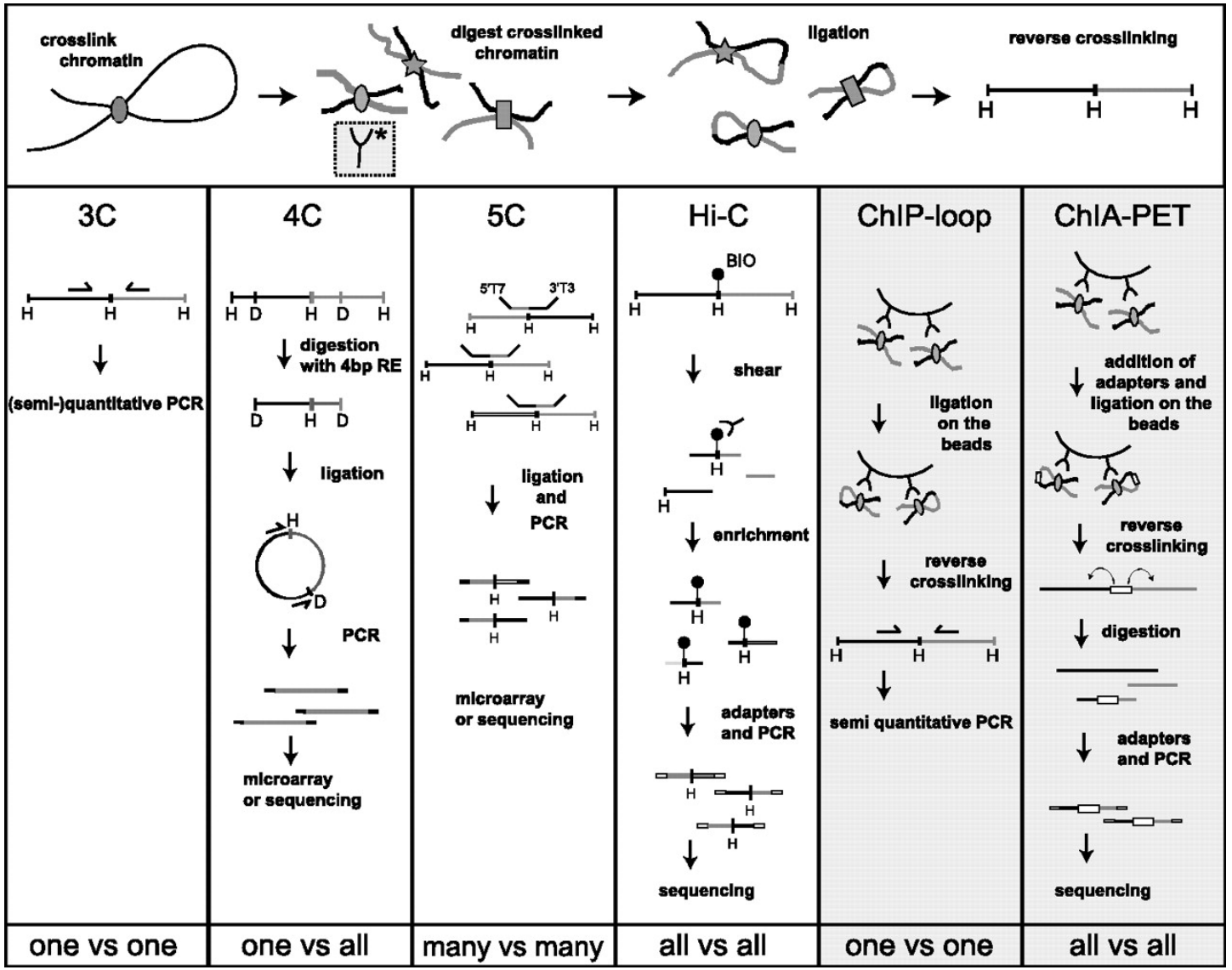
A



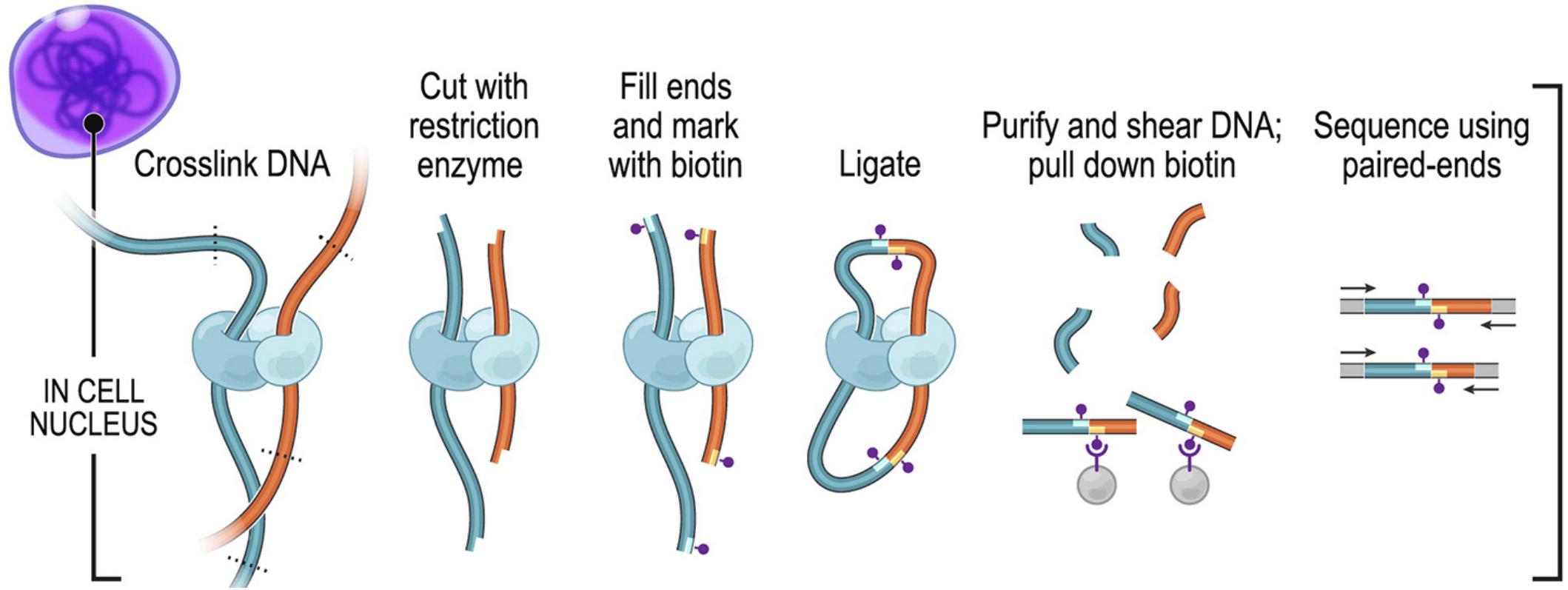
B

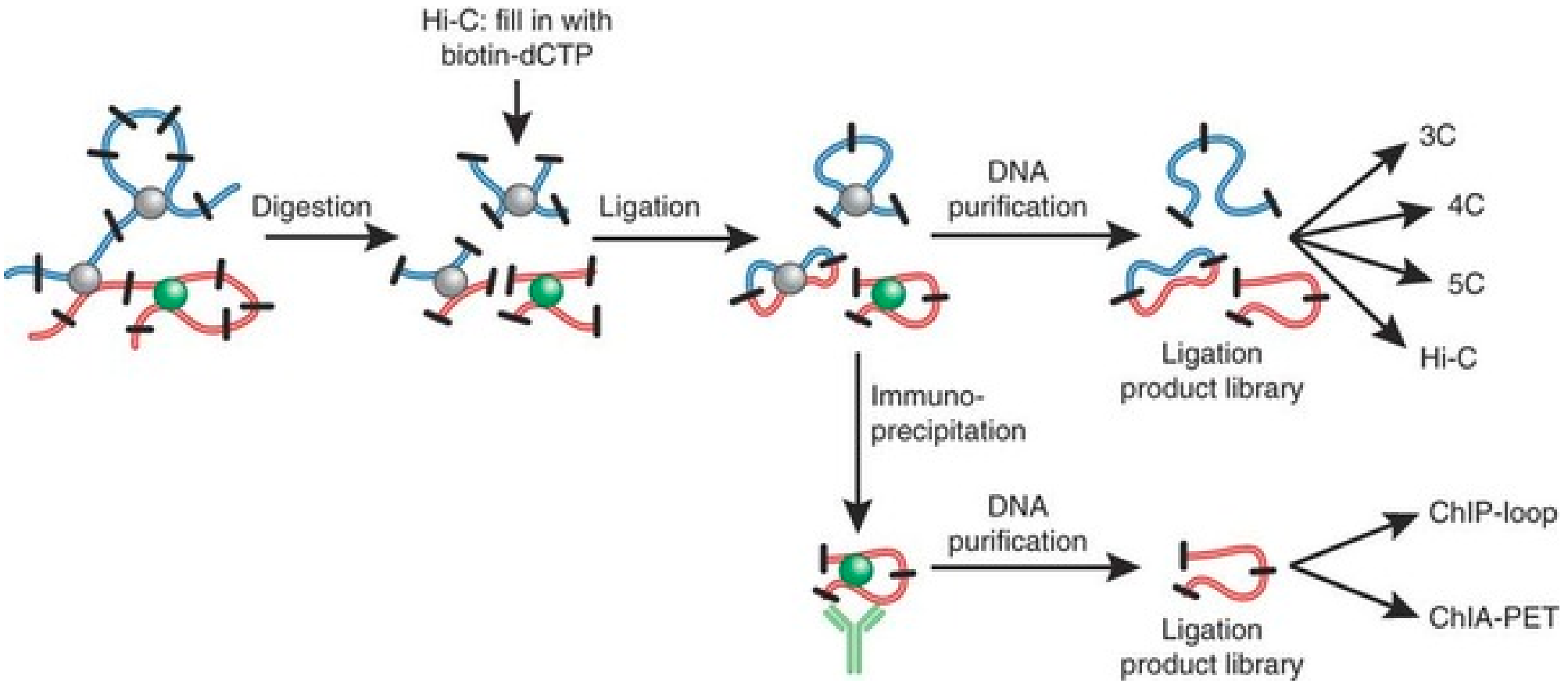


Overview of 3C-derived methods.



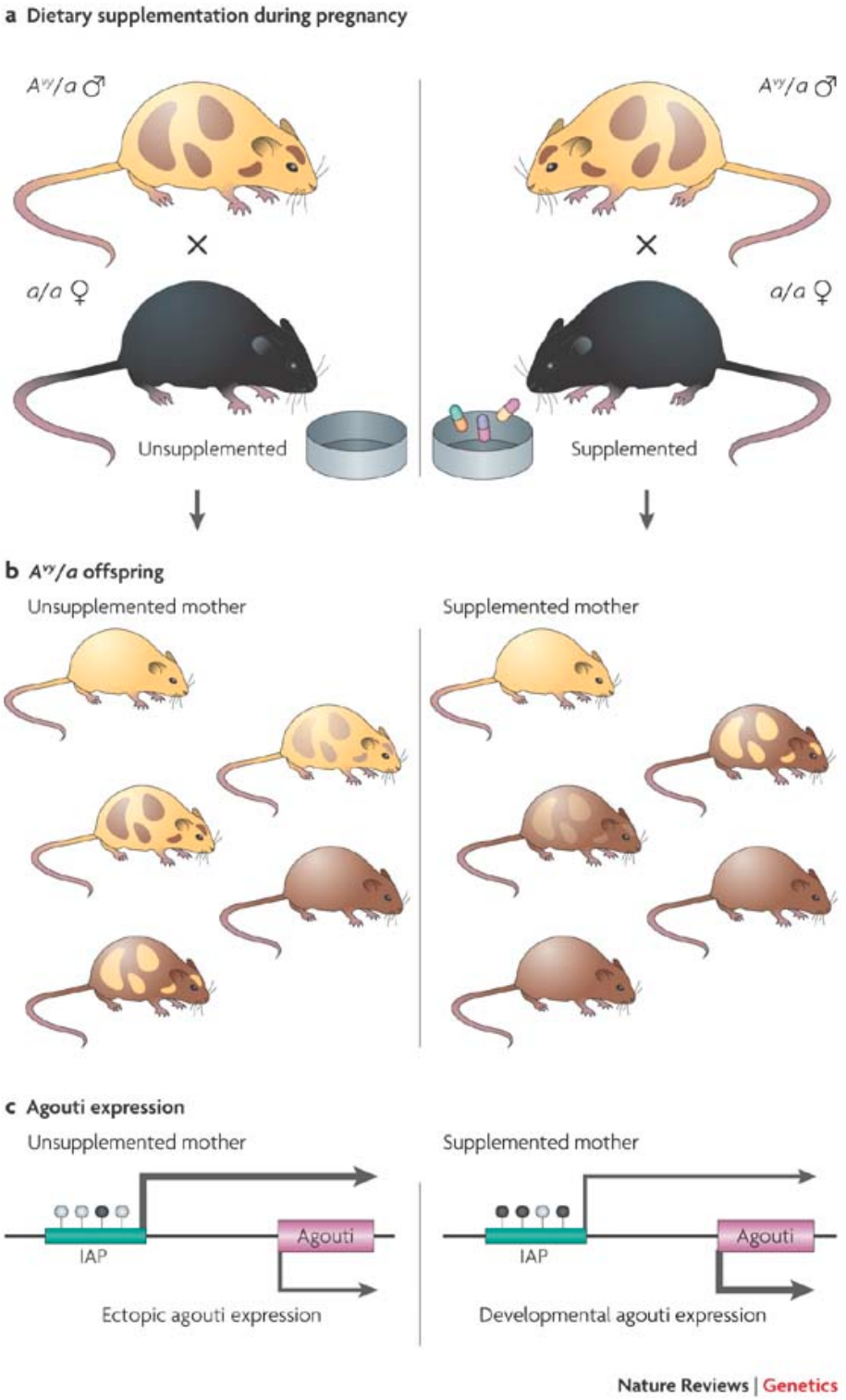
Elzo de Wit, and Wouter de Laat Genes Dev. 2012;26:11-24


Crosslink DNA
IN CELL NUCLEUS
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends


Hi-C: fill in with
biotin-dCTP
Digestion
Ligation
DNA
purification
3C
4C
5C
Hi-C
Ligation
product library
Immuno-
precipitation
DNA
purification
ChIP-loop
ChIA-PET
Ligation
product library

DNA Methylation

- DNA methylation and hydroxymethylation are involved in development, X-chromosome inactivation, cell differentiation, tissue-specific gene expression, plant epigenetic variation, imprinting, cancers, and diseases.
- Methylation usually occurs at the 5’ position of cytosines and plays a crucial role in gene regulation and chromatin remodeling.



- The active agouti gene in mice codes for yellow coat color. When pregnant mice with the active agouti gene are fed a diet rich in methyl donors, the offspring are born with the agouti gene turned off. This effect has been used as an epigenetic biosensor for nutritional and environmental alterations on the fetal epigenome.

Environmental epigenomics and disease susceptibility

Randy L. Jirtle and Michael K. Skinner

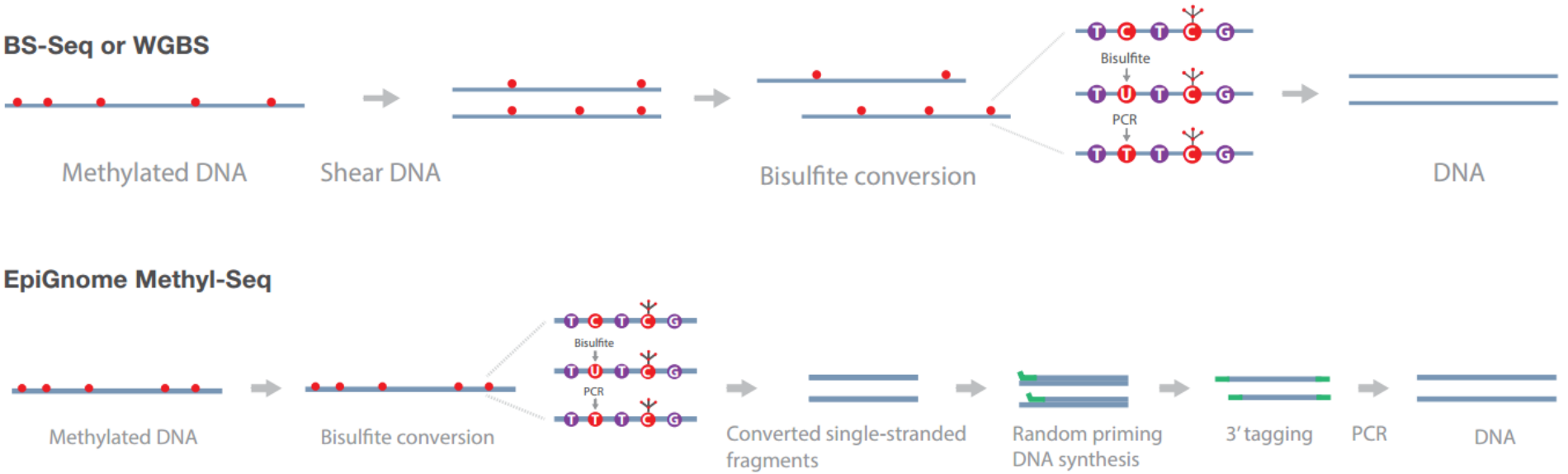
Nature Reviews Genetics 8, 253-262 (April 2007)

doi:10.1038/nrg2045

- Most cytosine methylation occurs on cytosines located near guanines, called CpG sites. These CpG sites are often located upstream of promoters, or within the gene body. CpG islands are defined as regions that are greater than 500 bp in length with greater than 55% GC and an expected/observed CpG ratio of > 0.65.
- While cytosine methylation (5mC) is known as a silencing mark that represses genes, cytosine hydroxymethylation (5hmC) is shown to be an activating mark that promotes gene expression and is a proposed intermediate in the DNA demethylation pathway. Similar to 5mC, 5hmC is involved during development, cancers, cell differentiation, and diseases.
- 5mC and/or 5hmC can be a diagnostic tool to help identify the effects of nutrition, carcinogens, and environmental factors in relation to diseases.
- The impact of these modifications on gene regulation depends on their locations within the genome. It is therefore important to determine the exact position of the modified bases.

BISULFITE SEQUENCING (BS-SEQ)

- Bisulfite sequencing (BS-Seq) or whole-genome bisulfite sequencing (WGBS) is a well-established protocol to detect methylated cytosines in genomic DNA.
- In this method, genomic DNA is treated with sodium bisulfite and then sequenced, providing single-base resolution of methylated cytosines in the genome. Upon bisulfite treatment, unmethylated cytosines are deaminated to uracils which, upon sequencing, are converted to thymidines.
- Simultaneously, methylated cytosines resist deamination and are read as cytosines. The location of the methylated cytosines can then be determined by comparing treated and untreated sequences.



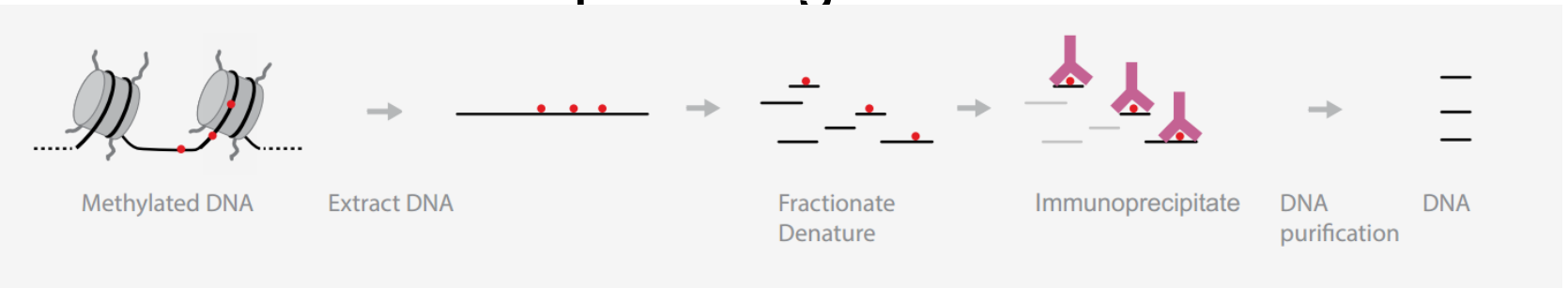
- Bisulfite converts unmethylated cytosines to thymidines, reducing sequence complexity, which can make it difficult to create alignments

- SNPs where a cytosine is converted to thymidine will be missed upon bisulfite conversion

- Bisulfite conversion does not distinguish between 5mC and 5hmC

METHYLATED DNA IMMUNOPRECIPITATION SEQUENCING (MEDIP-SEQ)

- Methylated DNA immunoprecipitation sequencing (MeDIP-Seq) is commonly used to study 5mC or 5hmC modification.
- Specific antibodies can be used to study cytosine modifications. If using 5mC-specific antibodies, methylated DNA is isolated from genomic DNA via immunoprecipitation.
- Anti-5mC antibodies are incubated with fragmented genomic DNA and precipitated, followed by DNA purification and sequencing.



Sequence Rearrangements

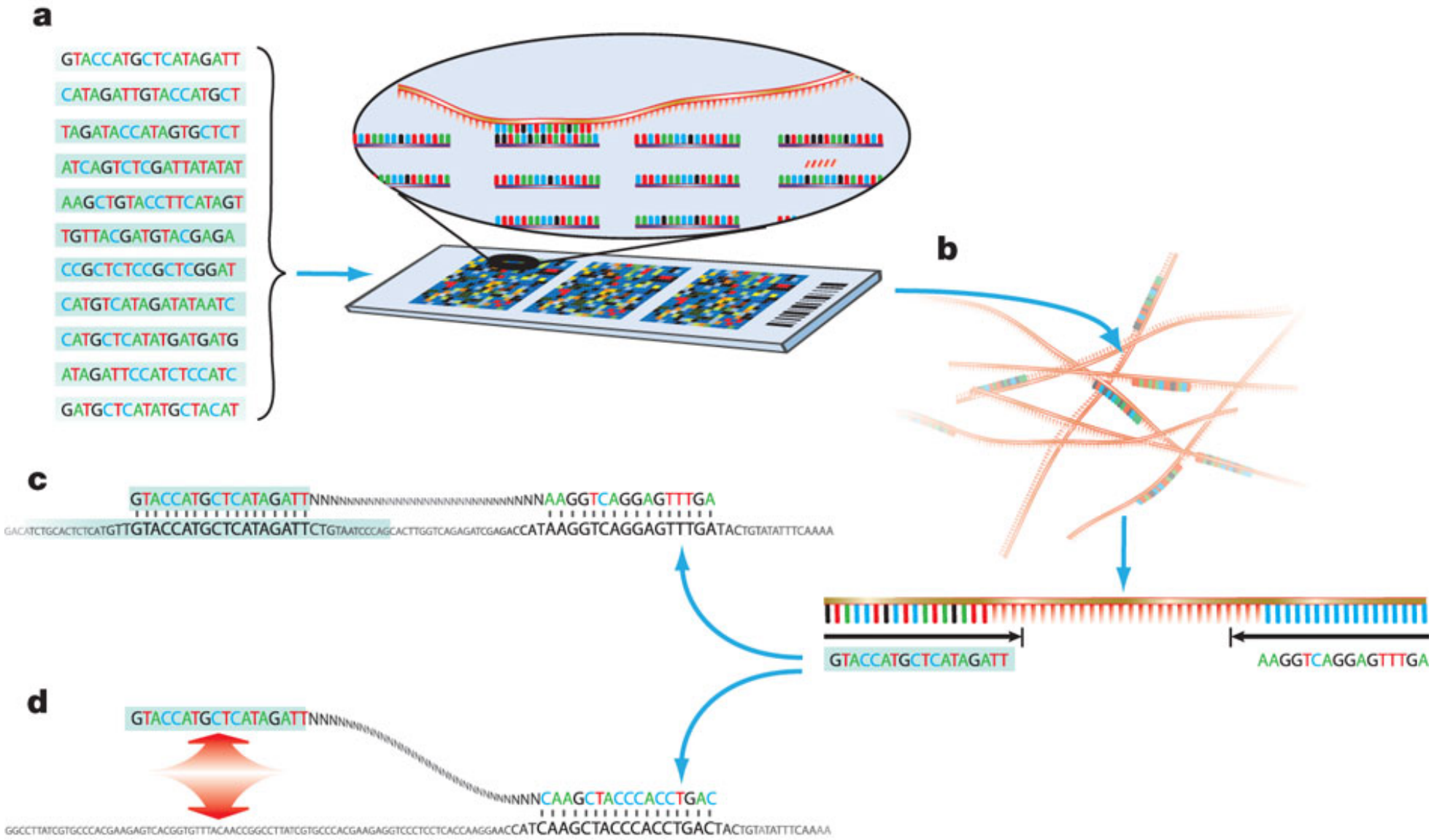
A growing body of evidence suggests that somatic genomic rearrangements, such as retrotransposition and copy number variants (CNVs), are relatively common in healthy individuals.

Cancer genomes are also known to contain numerous complex rearrangements. While many of these rearrangements can be detected during routine next-generation sequencing, specific techniques are available to study rearrangements such as transposable elements.

Transposable genetic elements (TEs) comprise a vast array of DNA sequences with the ability to move to new sites in genomes either directly by a cut-and-paste mechanism (transposons) or indirectly through an RNA intermediate (retrotransposons).

TEs make up about 66-69% of the human genome and play roles in ageing, cancers, brain, development, embryogenesis, and phenotypic variation in populations and evolution. TEs played a major role in dynamic arrangement of the sex determining region over evolution, giving us distinct X and Y chromosomes.

Overall RC-seq methodology.



JK Baillie *et al. Nature* **000**, 1-4 (2011) doi:10.1038/nature10531

nature