The CHOZN® Genomics Platform Facilitates Cell Line Engineering and Characterization

David Razafsky[1], Jiajian Liu[2], Rajagopalan Lakshminarasimhan[3],
Rahul Lal[3], Valeria Zanda[4], Fabio La Neve[4], Fei Zhong[2], Trissa Borgschulte[1]

1 Process Solutions, **MilliporeSigma**, St. Louis, USA; 2 Bioinformatics, **MilliporeSigma**, St. Louis, USA; 3 Bioinformatics, **Merck KGaA, Darmstadt, Germany**, Bangalore, India; 4 Advanced Sequencing Technologies, **Merck KGaA, Darmstadt, Germany**, Ivrea, Italy

Introduction

Chinese Hamster Ovary (CHO) cells have been the predominant cell line used for therapeutic protein production in the biopharmaceutical industry. Over the past several decades, significant improvements have been made to biotherapeutic production processes, resulting in enhanced protein production, better control of protein quality attributes and ultimately safer products for patients. Despite these advances, there is increasing pressure to reduce the cost of goods associated with biotherapeutic protein production. Genetic engineering offers the opportunity to enhance CHO cell line performance and reduce protein production costs. However, to take full advantage of new genetic engineering technologies, significant improvements in our understanding of the CHO genome and transcriptome are required.

Past efforts to sequence the CHO genome have been instrumental in improving our ability to genetically engineer CHO cells and have provided a better understanding of the underlying biology associated with industrial-scale protein production. While the second generation sequencing technologies used in these efforts remain a staple of any genome sequencing project, their short-read nature results in significant assembly gaps that hamper our ability to perform complex genetic- screening and engineering. In order for the biomanufacturing community to: 1) gain a more comprehensive view of the underlying mechanisms associated with favorable cellular phenotypes, 2) utilize cutting edge genetic screening strategies to enhance cell performance and 3) provide in-depth genetic characterization of manufacturing clones, a more contiguous genome assembly and more comprehensive gene annotation is required.

In an effort to improve the genomic and transcriptomic tools available, we have performed a *de novo* assembly and annotation of the CHOZN® GS$^{-/-}$ genome and transcriptome utilizing second generation (Illumina®) sequencing, third generation long-read (PacBio® Sequel) sequencing as well as ultra-long range interaction mapping via sequencing of Hi-C and CHiCAGO® libraries. The CHOZN® GS$^{-/-}$ hybrid assembly and annotation is available via a user-friendly, web-based interface that provides an unprecedented ability to characterize and genetically examine the host cell line and associated manufacturing clones and will play an important role in producing the next-generation of biomanufacturing host cell lines.

CHOZN® GS$^{-/-}$ Sequencing

High molecular weight genomic DNA (gDNA) was isolated from low passage CHOZN® GS$^{-/-}$ cells in the exponential growth phase. gDNA was used to prepare eight Illumina® libraries using the Nextera and TruSeq kits with insert sizes ranging from 430bp to 10.1kb. Libraries were sequenced on a MiSeq (PE300) and/or HiSeq2500 (PE100). Third party vendors prepared long-read SMRTbell™, Hi-C and CHiCAGO® libraries which were sequenced on a PacBio® Sequel and Illumina® HiSeqX (PE150) respectively.

To sequence the transcriptome, RNA was extracted from low passage cells in the exponential growth phase and libraries were prepared using the TruSeq Stranded RNA Kit with Ribo-Zero Globin. Libraries were sequenced on an Illumina® NextSeq500 (PE75).

Platform	Number of Reads	Estimated Sequencing Depth	Read Quality (% ≥Q30)	N50 Read Length
Illumina® HiSeq2500	$2.1x10^9$	>145x	>88%	N/A
Illumina® MiSeq	$3.5x10^7$	>5x	>72%	N/A
PacBio® Sequel	$4.6x10^6$	>20x	100%*	>20kb
Illumina® NextSeq500†	$1.3x10^8$	N/A	>80%	N/A

Table 1: Summary of CHOZN® GS$^{-/-}$ genome and transcriptome data collection.

*PacBio® Sequel separates all reads with an accuracy ≥80% from lower accuracy reads, which were then excluded from further analysis.

†Transcriptome sequencing.

CHOZN® GS$^{-/-}$ Assembly

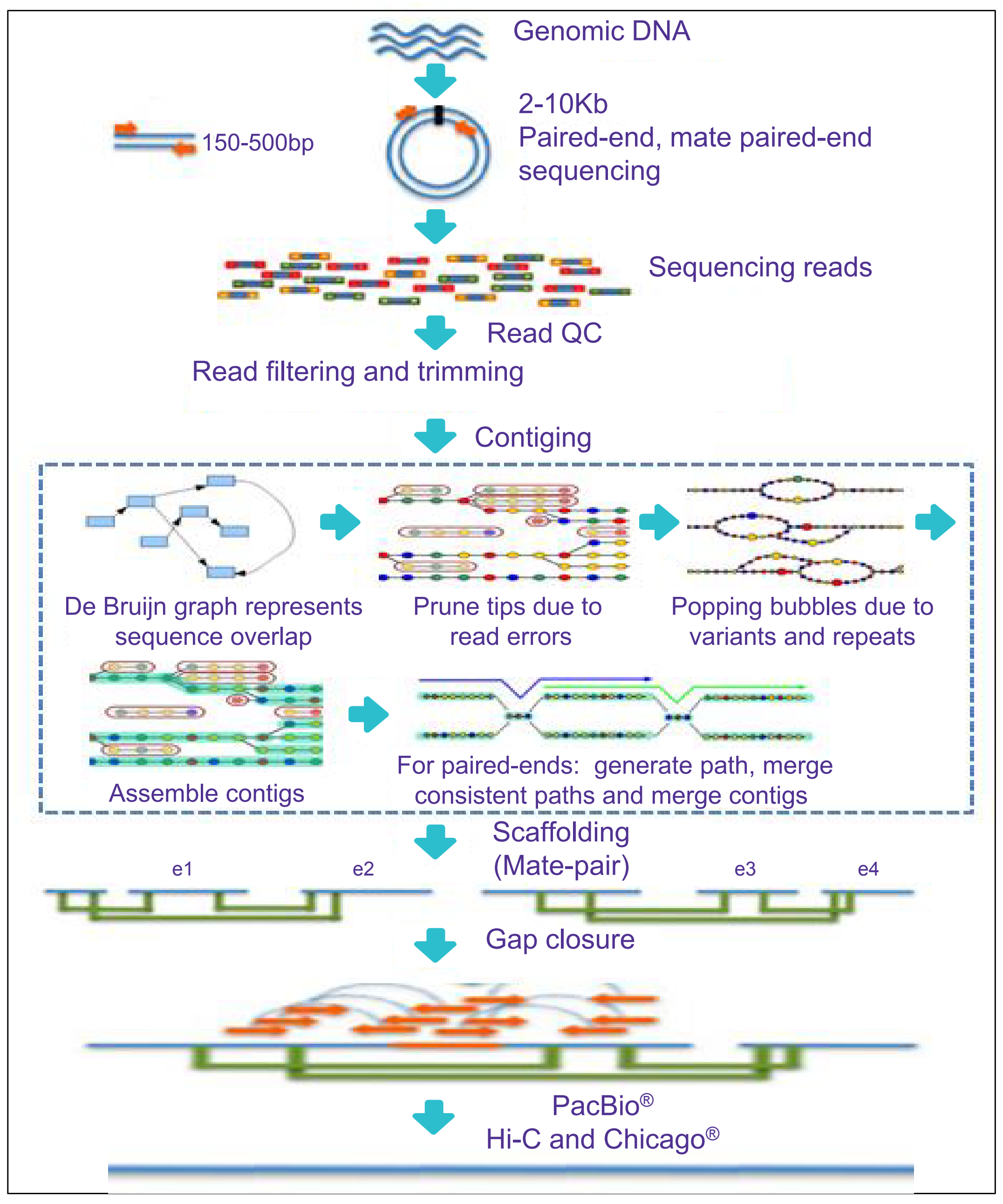

Genomic DNA
150-500bp
2-10Kb
Paired-end, mate paired-end sequencing
Sequencing reads
Read QC
Read filtering and trimming
Contiging
De Bruijn graph represents sequence overlap
Prune tips due to read errors
Popping bubbles due to variants and repeats
Assemble contigs
For paired-ends: generate path, merge consistent paths and merge contigs
Scaffolding (Mate-pair)
e1
e2
e3
e4
Gap closure
PacBio®
Hi-C and Chicago®

Figure 1: CHOZN® GS$^{-/-}$ assembly workflow. SOAP-denovo and Abyss were utilized to assemble the Illumina®-based paired-end and mate-paired reads from the libraries with different insert sizes, while Canu was used to assemble the genome from PacBio® reads. To further enhance the scaffold length and genome continuity, Hi-C and CHiCAGO® libraries were sequenced and incorporated into the final assembly by HiRise™.

CHOZN® GS$^{-/-}$ Gene Prediction and Annotation

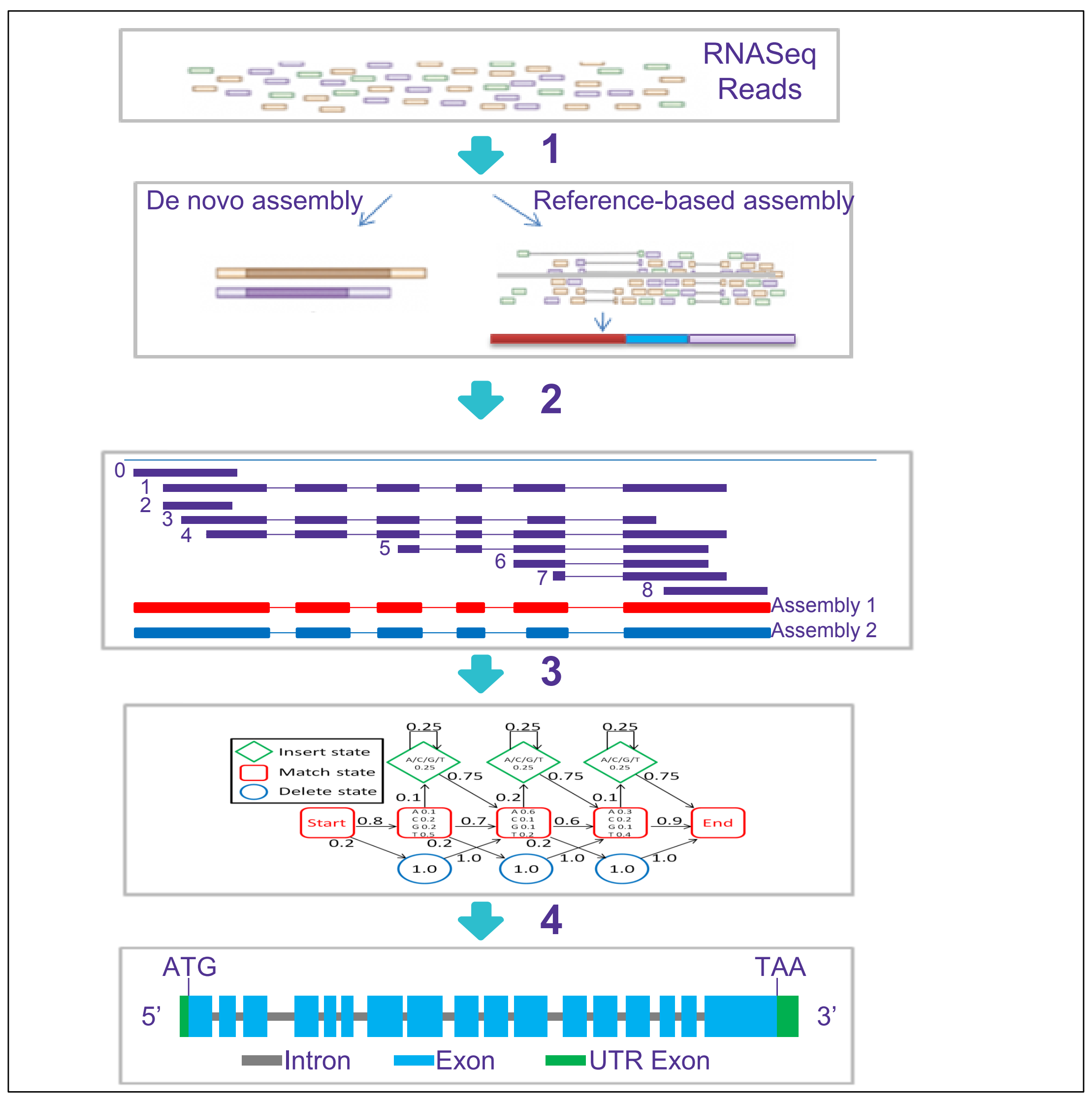


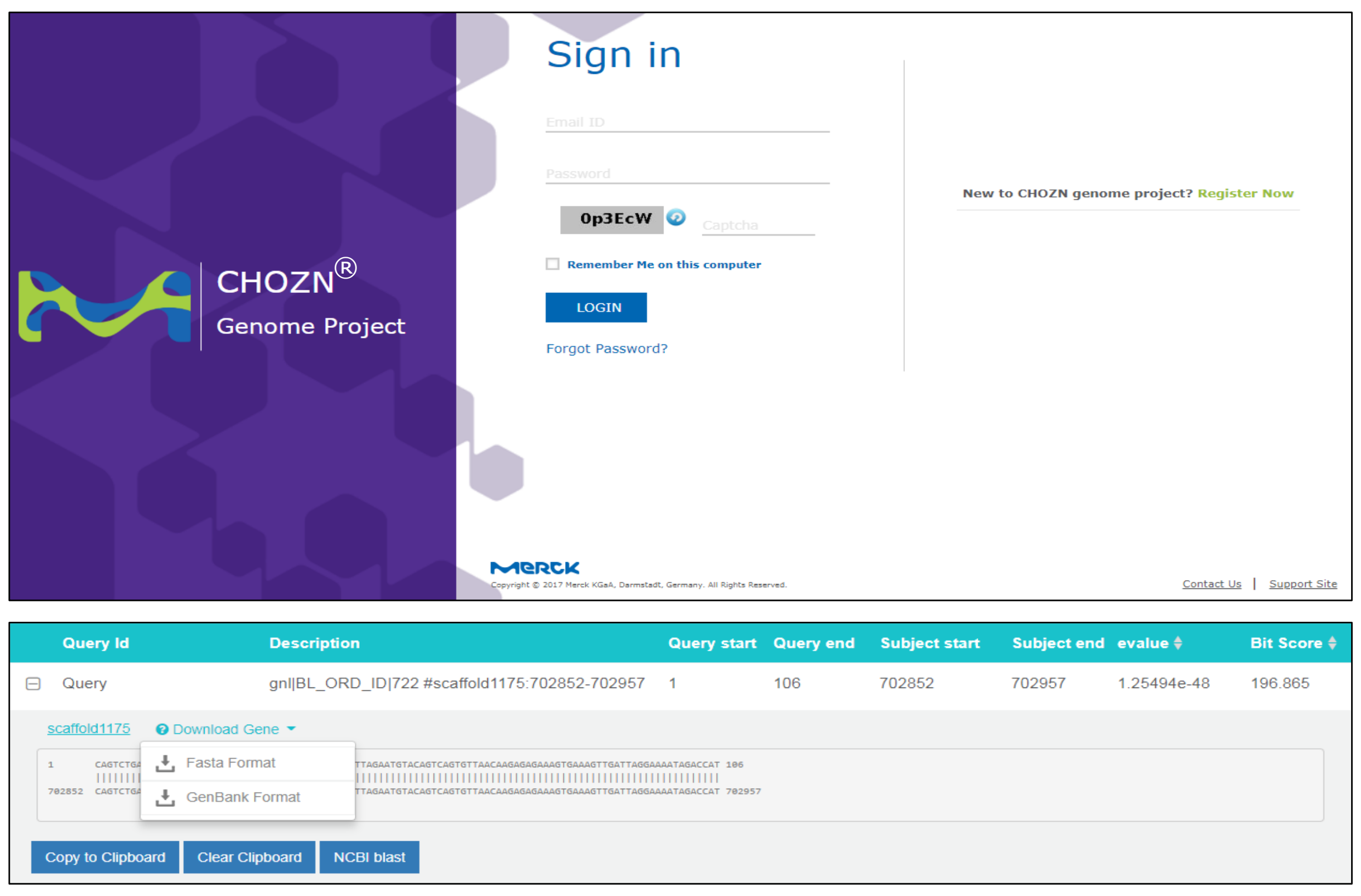
Figure 2: CHOZN® GS$^{-/-}$ annotation workflow. CHOZN® GS$^{-/-}$ gene prediction was primarily completed with Augustus gene-finding tool. A four step procedure was used for gene prediction in CHOZN® GS$^{-/-}$ cells. 1) After QC and filtering, RNASeq reads were assembled into transcripts. 2) A gene training set was constructed by aligning the assembled transcripts with genomic scaffolds. After clustering the aligned sequence groups, intron/exon gene structures were defined. 3) Gene-prediction HMM model parameters learned from the gene training sets were used to build gene models. 4) CHOZN® GS$^{-/-}$ gene prediction was completed using the defined model where assembled transcripts were used as hints. For the predicted genes, gene names/symbols were assigned based on their sequence orthology with mouse genes. Additional gene annotations were completed by identifying protein domains with Pfam, GO term and comparative genomic analysis with CHO-K1 and other related genomes.

CHOZN® GS$^{-/-}$ Assembly and Annotation Statistics

	CHOZN® GS$^{-/-}$ Assembly Statistics
Estimated Genome Size	2.66 Gb
Assembled Genome Size	2.48 Gb
Genome Assembled	93.2%
Maximum Scaffold Length	143 Mb
Scaffold N50	43.52 Mb
Scaffold L50	17
Scaffold N80	2.3 Mb
Scaffold L80	84
	CHOZN® GS$^{-/-}$ Annotation Statistics
Number of Predicted Genes	29,376
CHO-K1 Genes with CHOZN® GS$^{-/-}$ Counterpart	81.2%

Table 2: Summary of CHOZN® GS$^{-/-}$ assembly and annotation.

Secure Data Mining via a Web-based Interface



Query Id	Description	Query start	Query end	Subject start	Subject end	evalue	Bit Score
Query	gnl|BL_ORD_ID|722 #scaffold1175:702852-702957	1	106	702852	702957	1.25494e-48	196.865

```
1       CAGTCTGA  ...TTAGAATGTACAGTCAGTGTTAACAAGAGAGAAAGTGAAAGTTGATTAGGAAAATAGACCAT 106
702852  CAGTCTGA  ...TTAGAATGTACAGTCAGTGTTAACAAGAGAGAAAGTGAAAGTTGATTAGGAAAATAGACCAT 702957
```

Figure 3: CHOZN® GS$^{-/-}$ User Interface. The web-based interface allows users to quickly search the genome and transcriptome via BLAST or gene-name/symbol. Users can then download the corresponding sequence(s) in fasta or GenBank format.

Summary

CHO cells have been the cell line of choice in biologics production for more than 30 years. While significant improvements have been made in production processes, enhancements to the cell line have lagged behind due to the inadequate genomic resources that have been available. Here we provide a considerably improved CHO assembly and annotation as well as a web-based, user friendly, interface to ease data mining. These tools will allow CHOZN® GS$^{-/-}$ customers to: 1) plan complex genome engineering strategies, 2) design and utilize genome-wide screening tools, 3) characterize transgene integration events with increased precision and confidence and 4) better understand the underlying biology of favorable phenotypes.

The life science business of Merck operates as MilliporeSigma in the U.S. and Canada.

SigmaAldrich.com

SAFC®
Pharma & Biopharma Raw Material Solutions