GARVAN INSTITUTE

CONFIDENTIAL

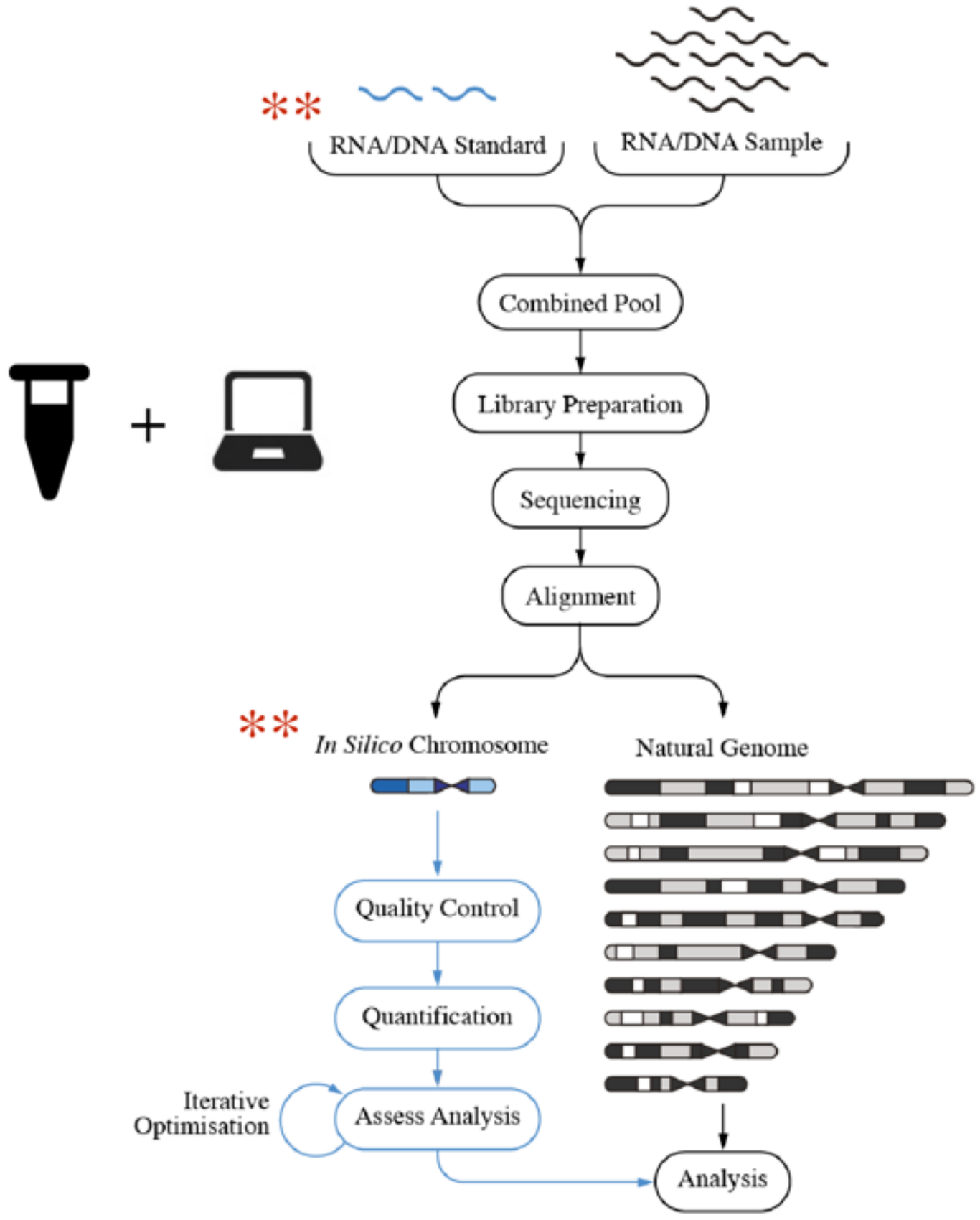
Spike-In Standards for Next Generation Sequencing

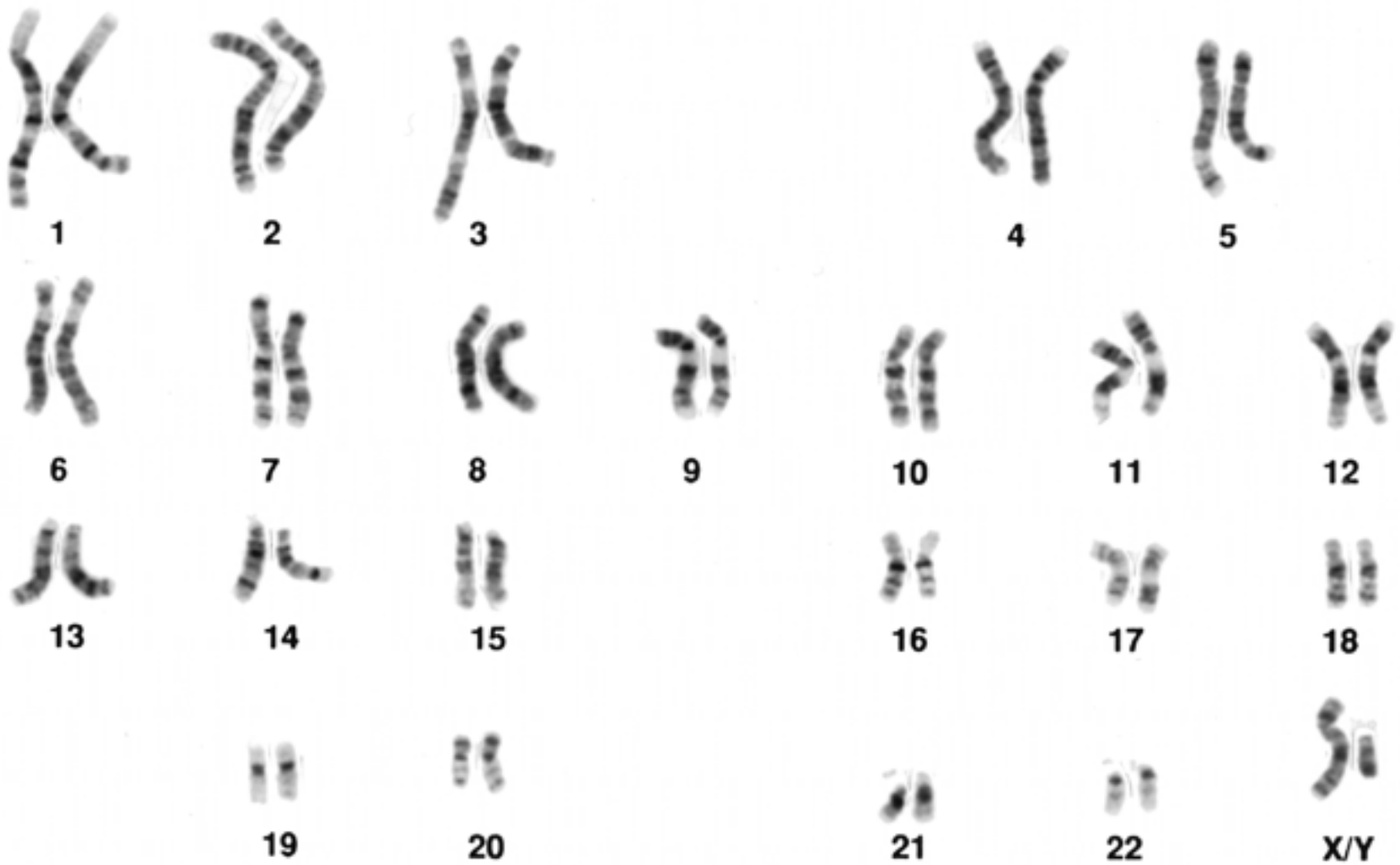
by Tim Mercer

Confidential Under Disclosure Agreement between

Garvan Institute of Medical Research and Roche NimbleGen Inc.

dated 20 June 2013 as amended by variation agreements dated 1 October 2013 and 10 February 2015.

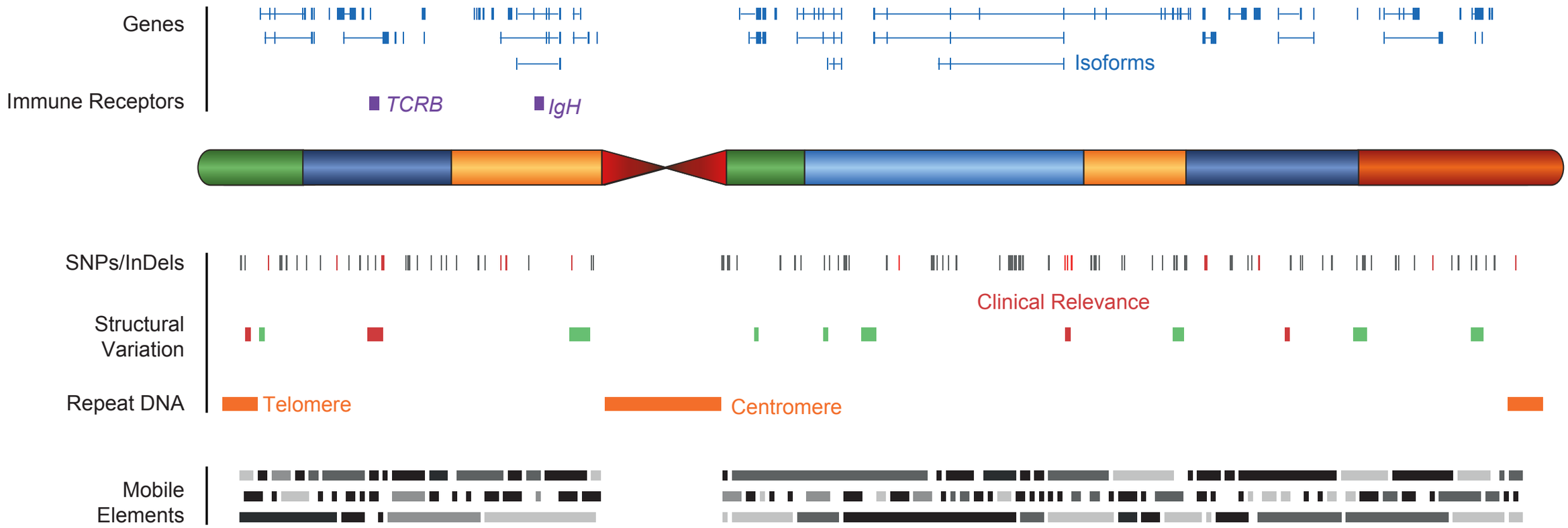




How to make an *in silico* chromosome.

1. Shuffle (retains nucleotide composition)
2. Invert (retains nucleotide and repeat composition)
3. Convert (retain GC and repeat composition)
4. Manual (artisanal for specific genetic features)

In Silico Chromosome.



No homology to natural sequences but encode the following features:

1. **Genes** (alternatively spliced into isoforms)
2. **Fusion Genes**
3. **Immune Receptors** (undergo V(D)J recombination and hypermutation)
4. **SNPs/Indel Genetic Variation** (homo- and heterozygous)
5. **Structural Variation**
6. **Repetitive DNA** (satellite DNA, centromeres, telomeres)
7. **Mobile elements** (SINE, LINE, LTR-like elements)

Conjoined Standards.

Standard abundance = concentration x copy number

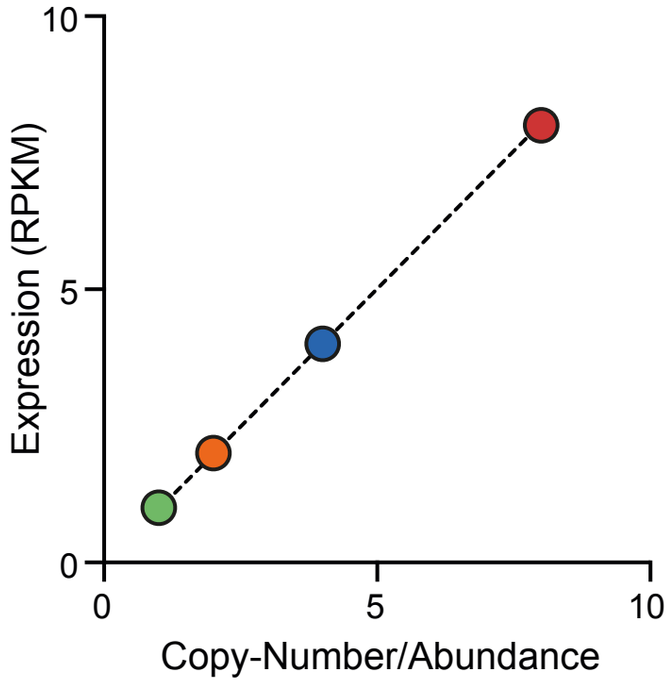
Concentration can establish a wide range in abundance between lowest and highest standards, but pipetting errors adds variation.

Individual standards of different abundance

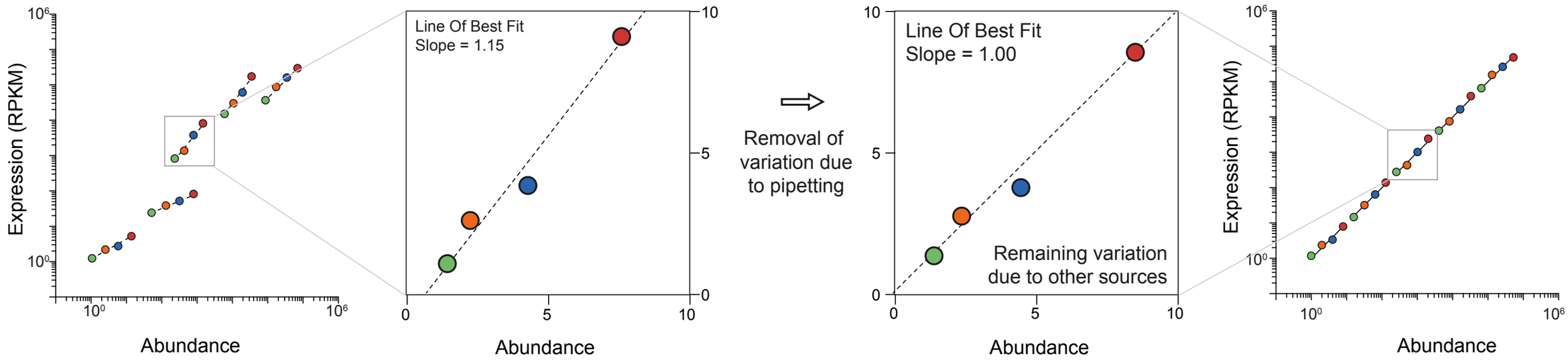
OR

Individual standards of different copy-number joined in single standard

Copy-number avoids pipetting errors, but has a limited range in abundance between lowest and highest standards.

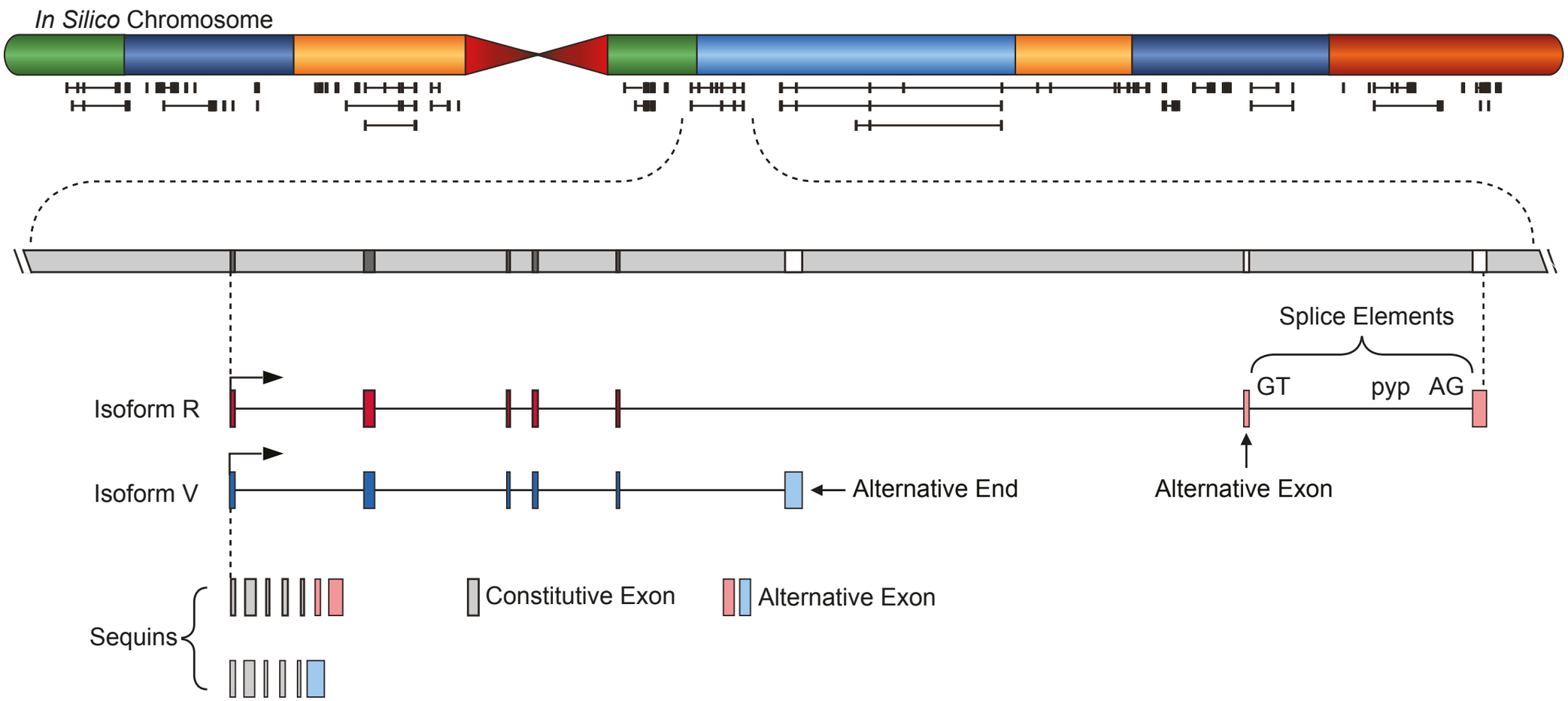


Using both concentration and copy-number to establish abundance can establish a wide range in abundance *and* remove pipetting variation.



Force the line of best fit to = 1 between individual copies on a single standard (dependent variation) removes errors due to pipetting, providing maximum quantitative accuracy between the standard in a mixture.

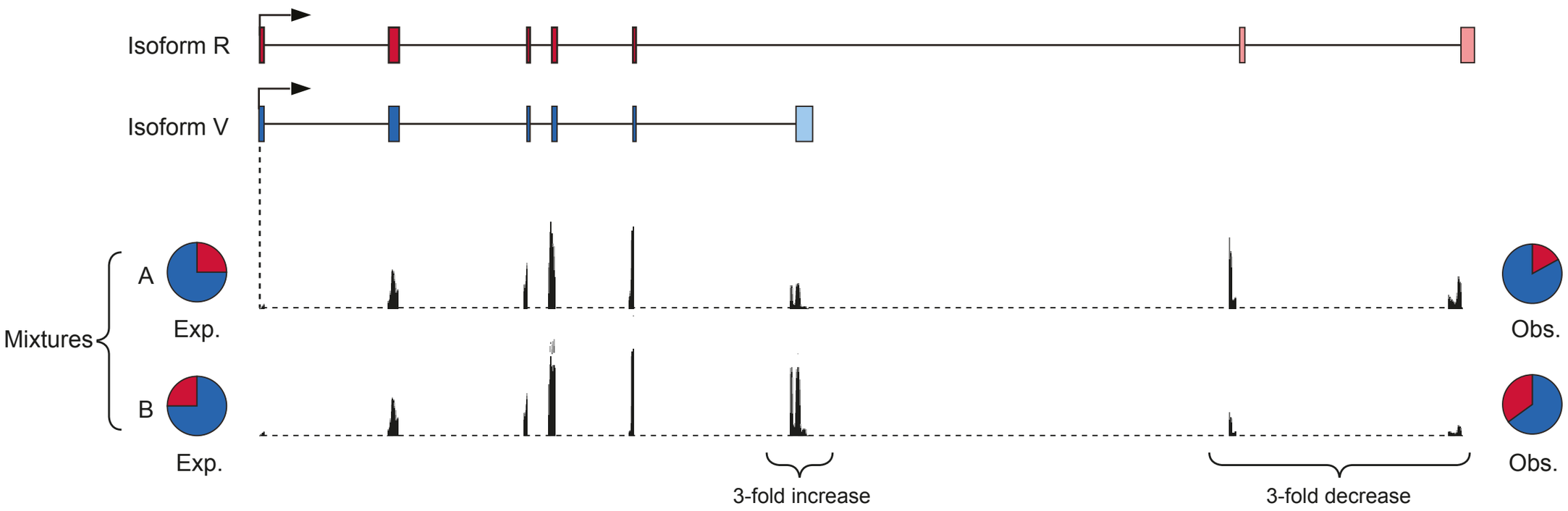
Artificial Gene loci.

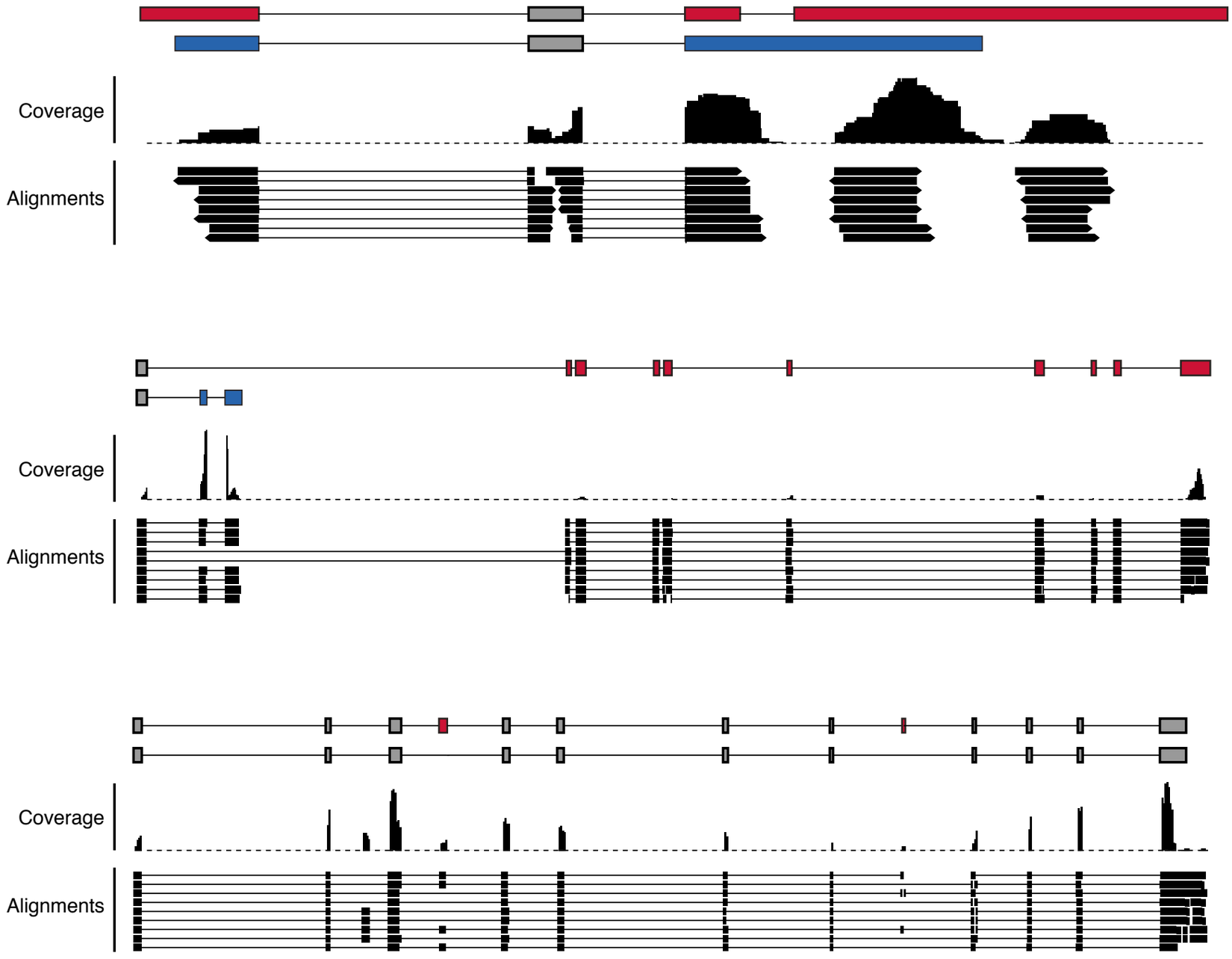


1. Currently have 66 genes comprising 142 isoforms.
2. Each mixture comprises 22 different concentration represented by genes encompassing 10^6 fold dynamic range.

Alternative Splicing.

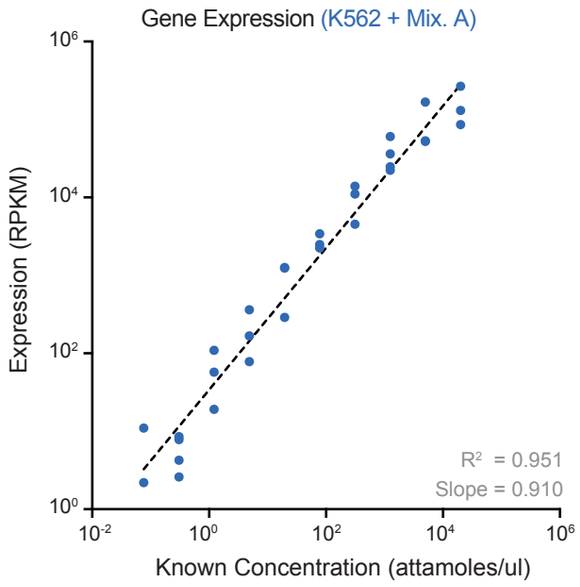
Varying the relative ratio of isoforms emulates quantitative changes in alternative splicing.

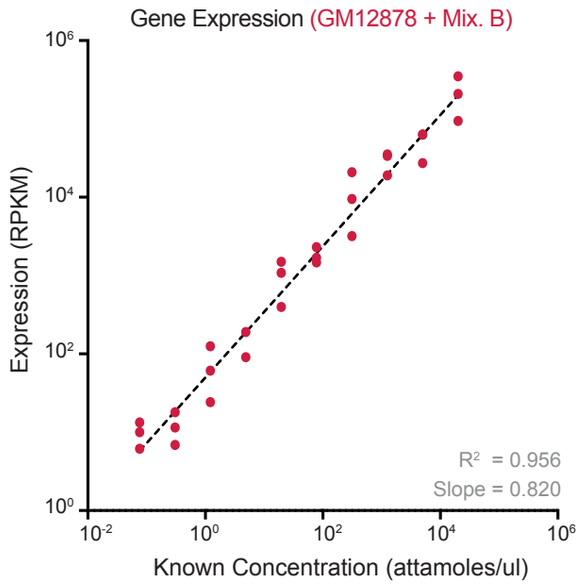


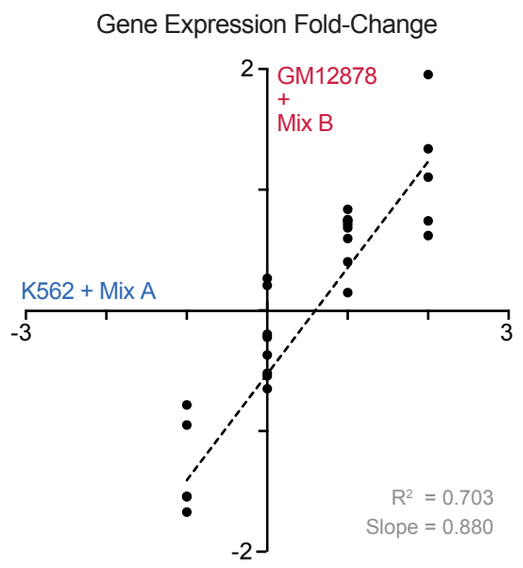


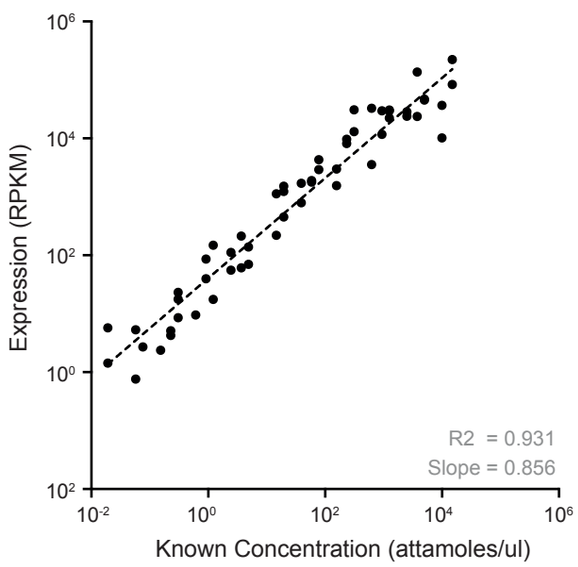
A representative range of alternative splicing events are emulated.

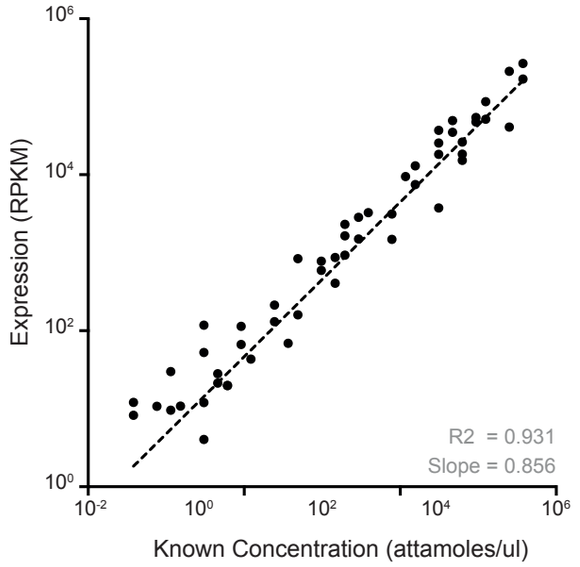
Differential Gene Expression.

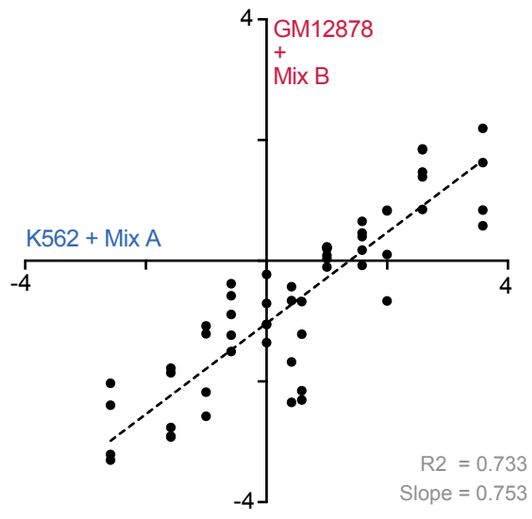








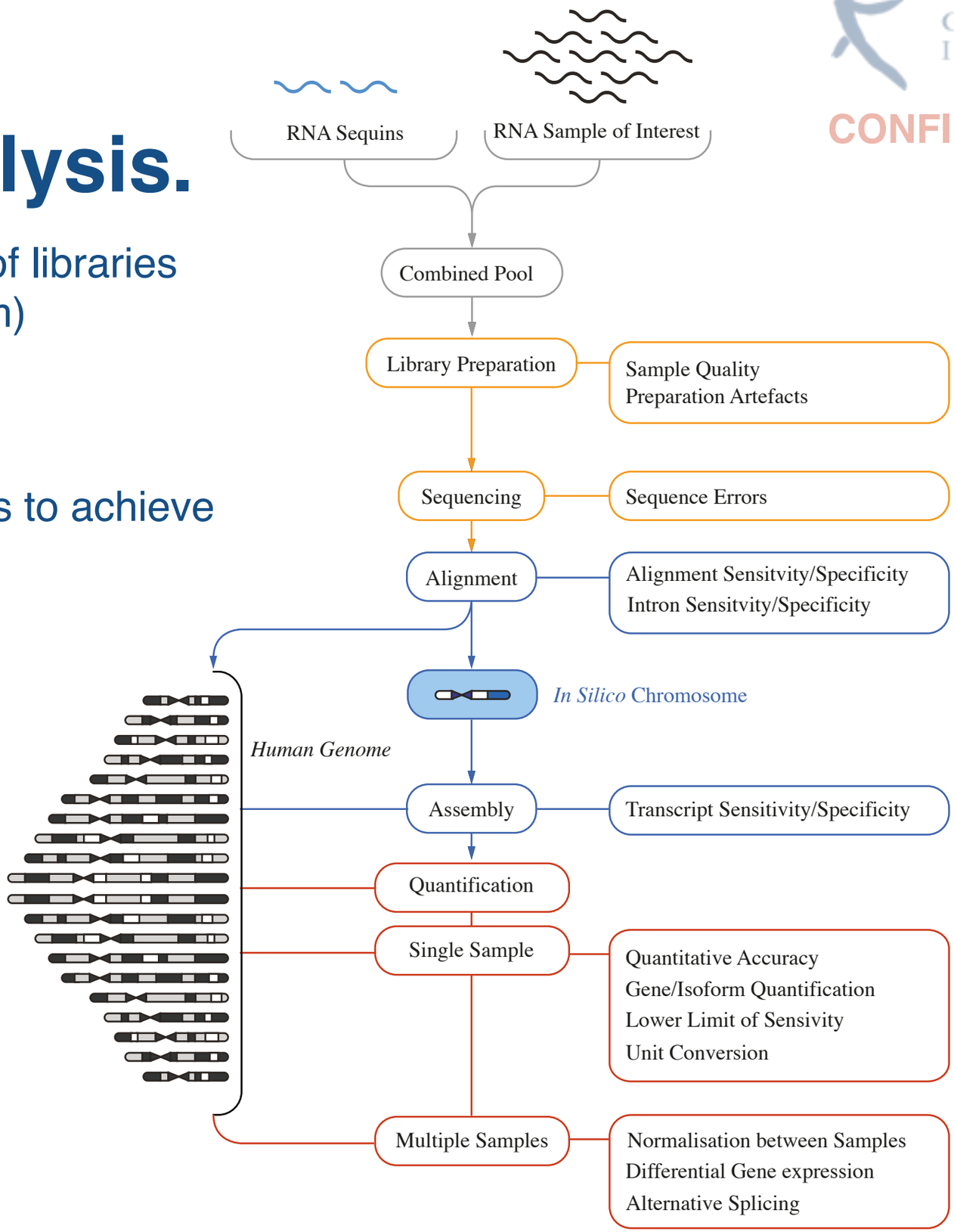




Adding different mixtures of standards to different samples emulates differential gene expression and alternative splicing.

Bioinformatic Analysis.

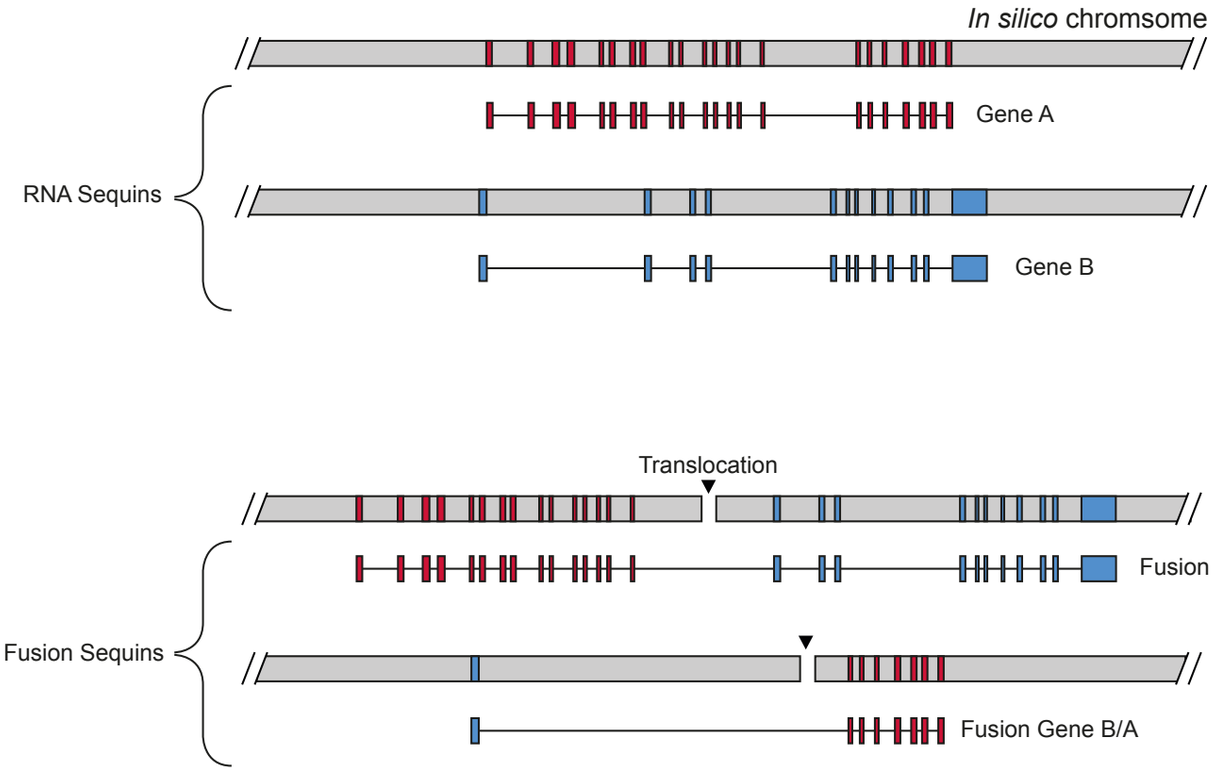
1. Provides automated assessment of libraries (.fastq) and alignments (.bam/.sam)
2. Enables quality control and rapid troubleshooting.
3. Enables real-time iterative analysis to achieve optimal results.

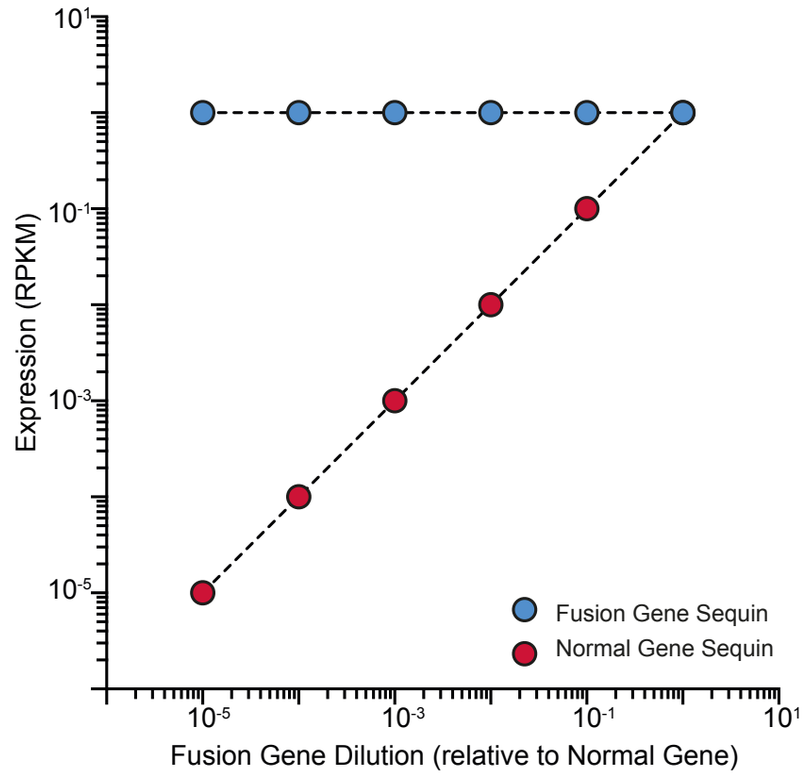


Assessment of RNA sequencing

RNA Sample	**None**		**K562**		**GM12878**		**Mouse Liver**		**Lung Normal**		**Lung Cancer**	
RNA Standards Mixture	**Flat**		**MixA**		**MixB**		**MixA**		**MixB**		**MixA**	
Reads to Genome	22,025,286.00		2,183,367.00		1,412,061.00		225,787.00		46,351.00		189,582.00	
Reads to *In Silico*	0.00		131,557,592.00		88,572,155.00		14,983,308.00		48,348,495.00		81,862,028.00	
Fraction Dilution	1.0000		0.0163		0.0157		0.0148		0.0010		0.0023	
Alignment	**Sn**	**Sp**	**Sn**	**Sp**	**Sn**	**Sp**	**Sn**	**Sp**	**Sn**	**Sp**	**Sn**	**Sp**
All Alignments	0.98	0.96	0.81	0.84	0.89	0.83	0.64	0.86	0.57	0.78	0.63	0.80
Spliced Alignments	0.99	0.93	0.86	0.85	0.92	0.82	0.72	0.94	0.61	0.84	0.70	0.84
Detection Limit	-	-	0.0796	-	0.0018	-	0.0181	-	0.0049	-	0.0112	-
Assembly												
Base level	86.30	99.50	0.71	0.95	0.75	0.95	0.56	0.97	0.47	0.92	0.51	0.92
Exon level	80.00	98.30	0.70	0.77	0.75	0.79	0.60	0.86	0.45	0.67	0.53	0.73
Intron level	79.20	100.00	0.72	0.95	0.77	0.98	0.63	0.98	0.49	0.92	0.56	0.94
Intron chain level	66.70	65.60	0.40	0.38	0.50	0.48	0.40	0.50	0.28	0.35	0.33	0.42
Locus level	71.90	59.00	0.53	0.21	0.59	0.23	0.47	0.33	0.38	0.18	0.41	0.19
Missed exons	16.10	-	0.22	-	0.17	-	0.35	-	0.48	-	0.42	-
Missed Introns	18.10	-	0.19	-	0.17	-	0.33	-	0.43	-	0.22	-
Missed Loci	6.20	-	0.22	-	0.13	-	0.34	-	0.34	-	0.25	-
Expression												
Gene Correlation	-	-	0.95	-	0.95	-	0.94	-	0.90	-	0.93	-
Slope	-	-	0.91	-	0.96	-	0.91	-	0.92	-	0.94	-
Y-int	-	-	1.53	-	1.66	-	1.45	-	1.73	-	1.38	-
Isoform Correlation	-	-	0.93	-	0.87	-	0.93	-	0.73	-	0.90	-
Slope	-	-	0.86	-	0.84	-	0.83	-	0.75	-	0.86	-
Y-int	-	-	-1.08	-	2.16	-	1.63	-	-1.08	-	1.53	-
Alt. Splicing Correlation	-	-	0.64	-	0.49	-	0.57	-	0.20	-	0.78	-
Slope	-	-	0.84	-	0.62	-	0.78	-	0.52	-	0.97	-

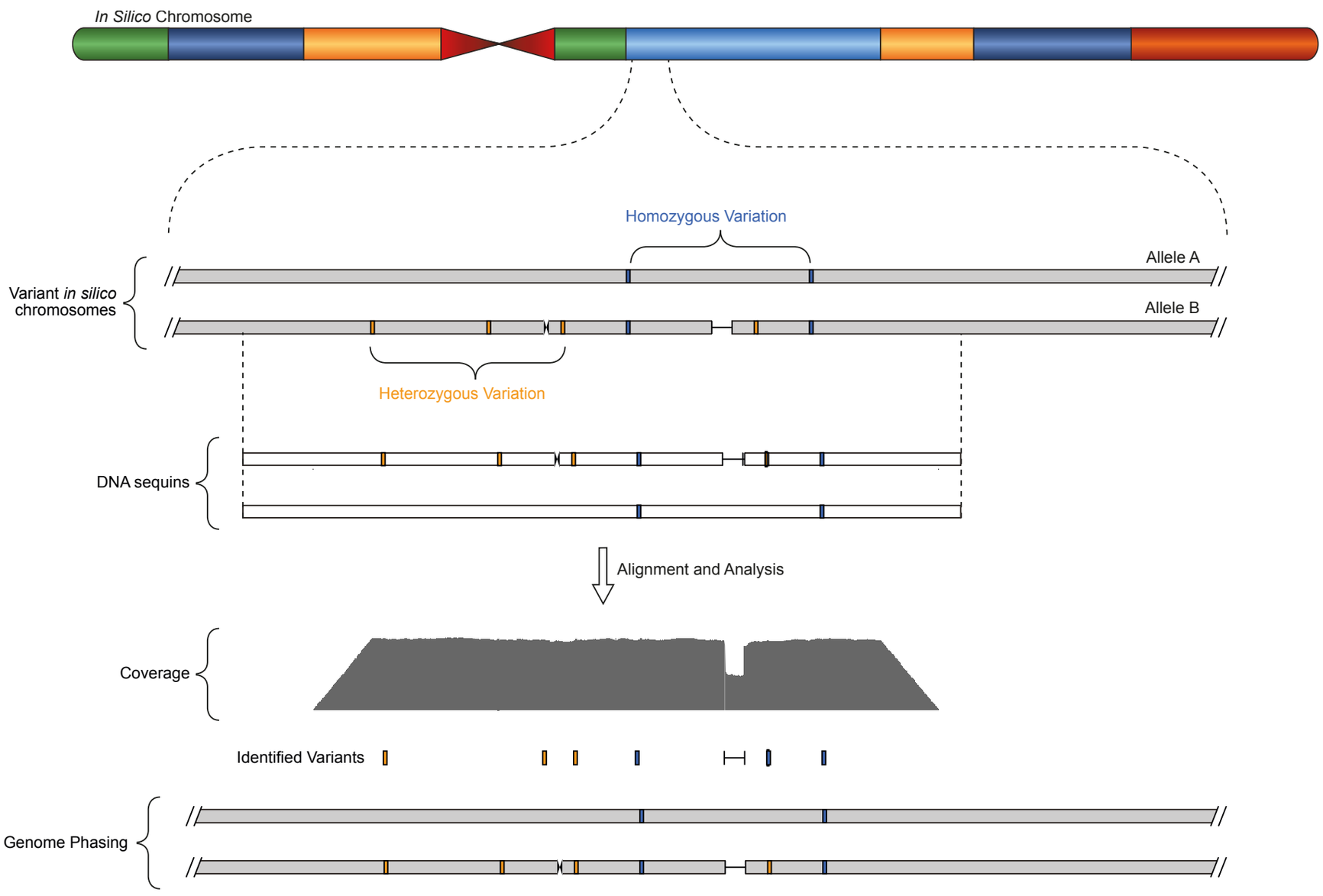
Fusion Genes.





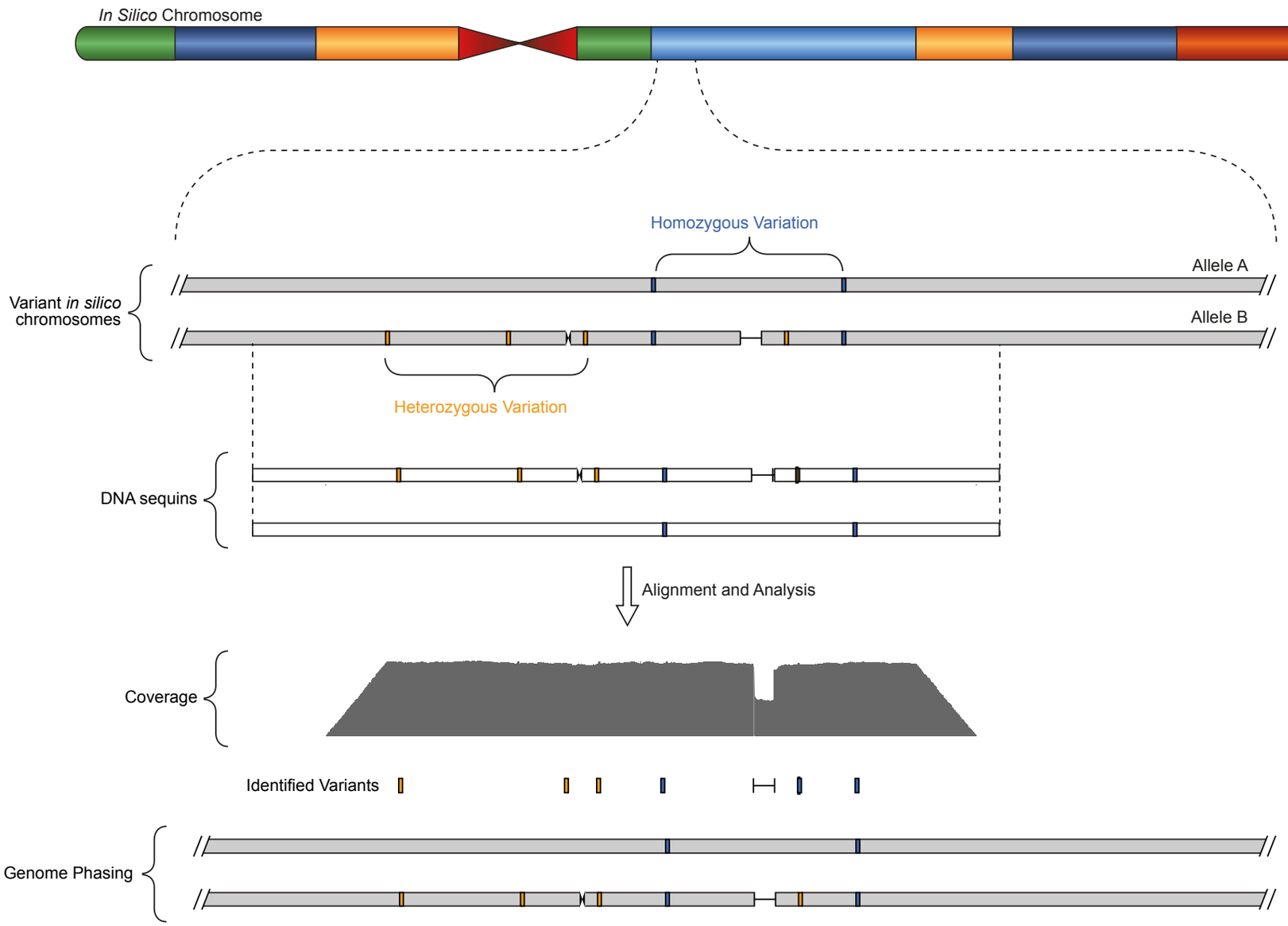
1. 12 Fusion events are each represented with 2 normal genes and 2 reciprocal fusion genes (48 RNA standards in total)
2. Assess detection and quantification of fusion genes and limit of sensitivity
3. Applications include cancer diagnosis and monitoring minimal residual disease

Genetic Variation.



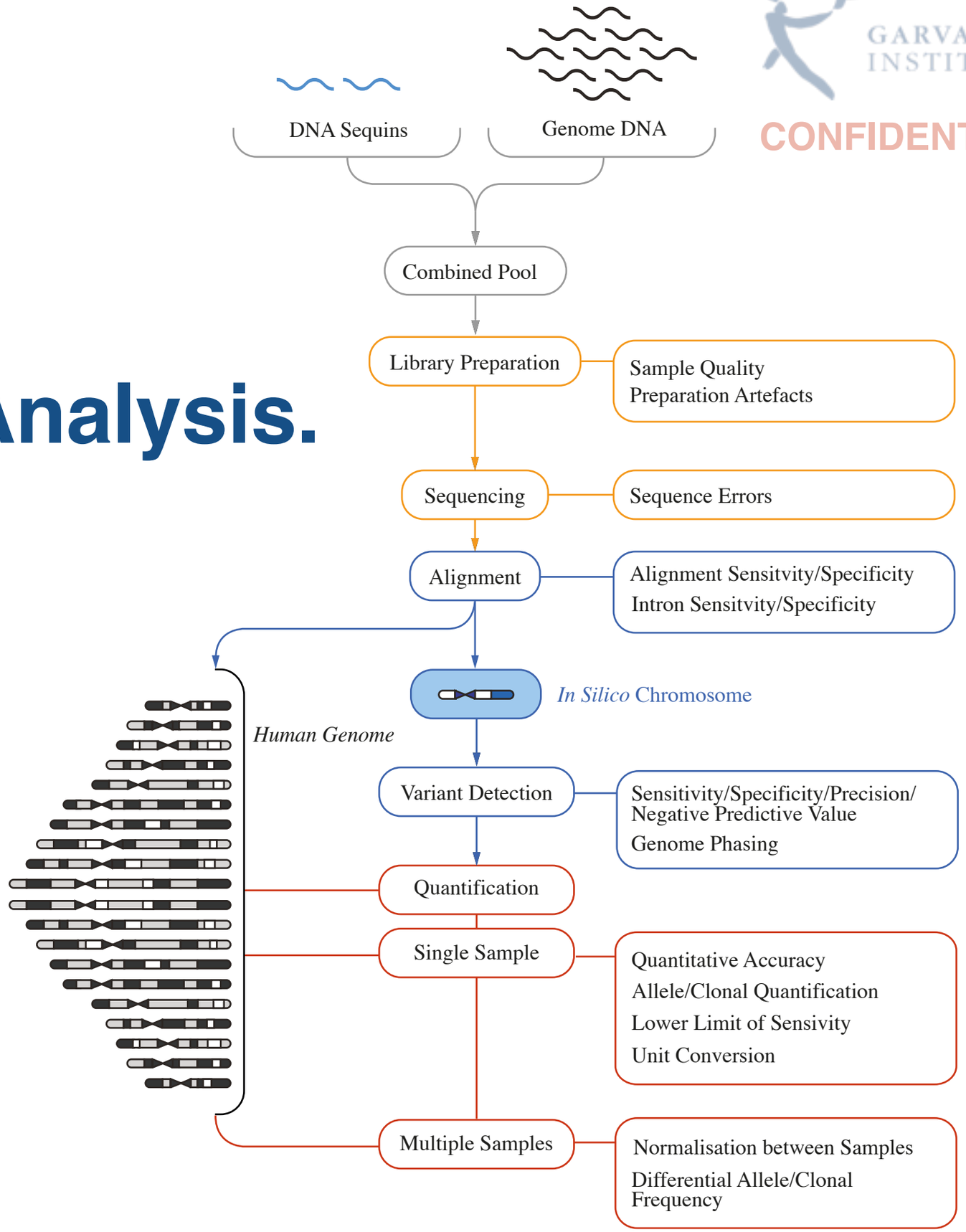
1. Representative human genetic variation is incorporated *in silico* chromosome variants
2. 66 DNA standard represent the reference or variant *in silico* chromosome sequences (132 DNA standards in total encompass 253 genetic variations
3. Genetic variants retain the local sequence contest up to 10nt flanking.

Genetic Variation.



1. Varying the ratio of reference and variant DNA standards emulates homozygous, heterozygous genotypes.
2. Heterogenous allele frequencies, such as when only a sub-population of sample harbours genotypes, are represented

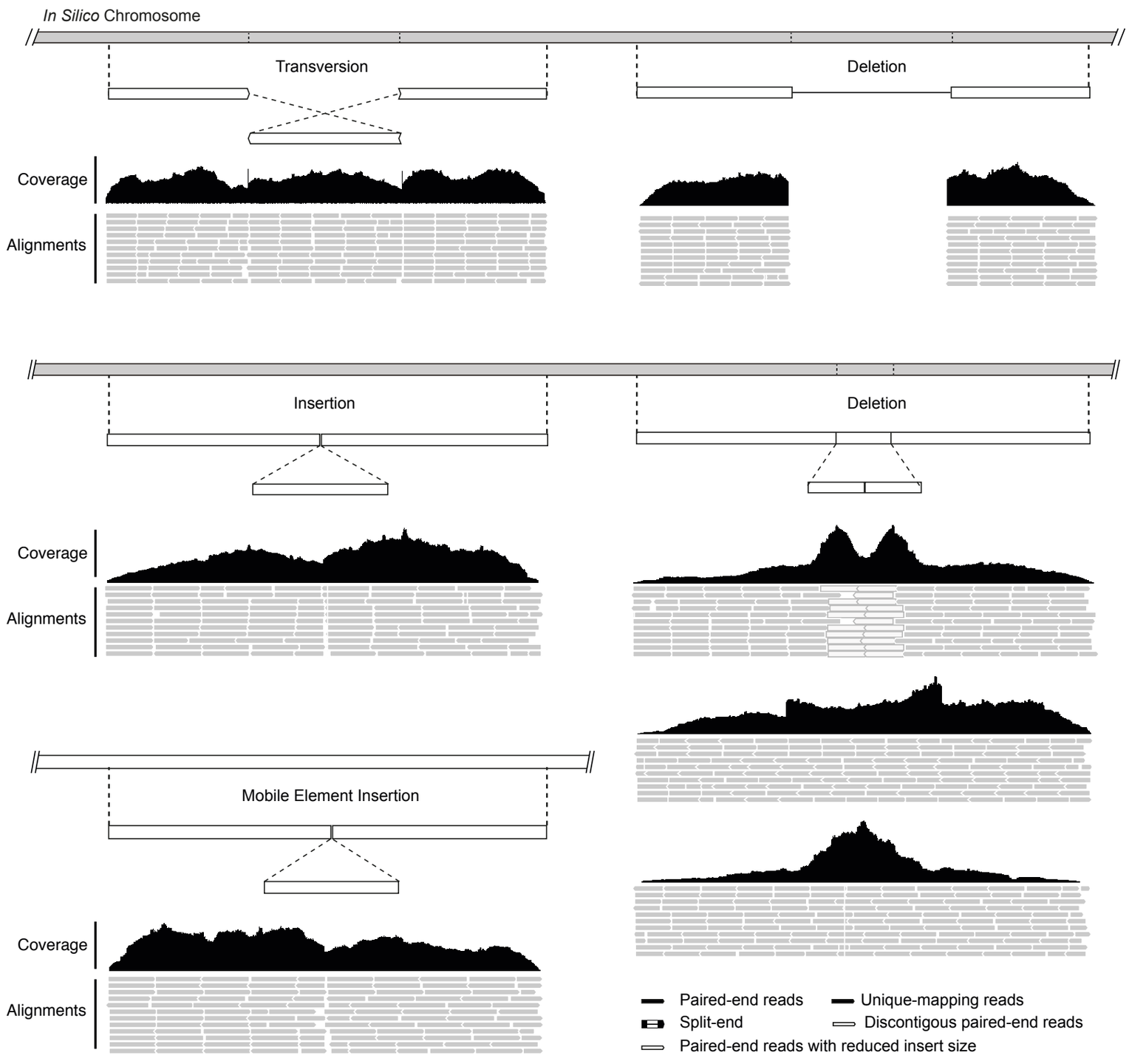
Bioinformatic Analysis.



Assessment of whole genome sequencing.

Genome DNA Standards	None Flat	GM12878 MixA	Mouse Liver MixA	Lung Normal MixA	Lung Cancer MixB
Reads to Genome	1,000,000,000	458,521,347	11,799,567.00	221,843,673.00	233,622,706.00
Reads to *In Silico*	1,000,000	2,029,597	70,492.00	1,532,753.00	2,022,901.00
Fraction Dilution	0.0100	0.0040	0.0059	0.0086	0.0069
Alignment					
Sensitivity	0.990	0.849	0.758	0.780	0.820
Specificity	0.860	0.961	0.848	0.970	0.960
Detection Limit	-	0.004	0.012	0.002	0.002
Variation					
Covered	0.996	0.880	0.720	0.890	0.910
Detection Sensitivity	0.986	0.650	0.570	0.730	0.780
Efficiency	0.990	0.730	0.680	0.780	0.810
Specificity	0.931	0.570	0.410	0.650	0.710
Aundance					
Correlation	-	0.949	0.967	0.980	0.990
Slope	-	0.910	0.950	1.030	1.010

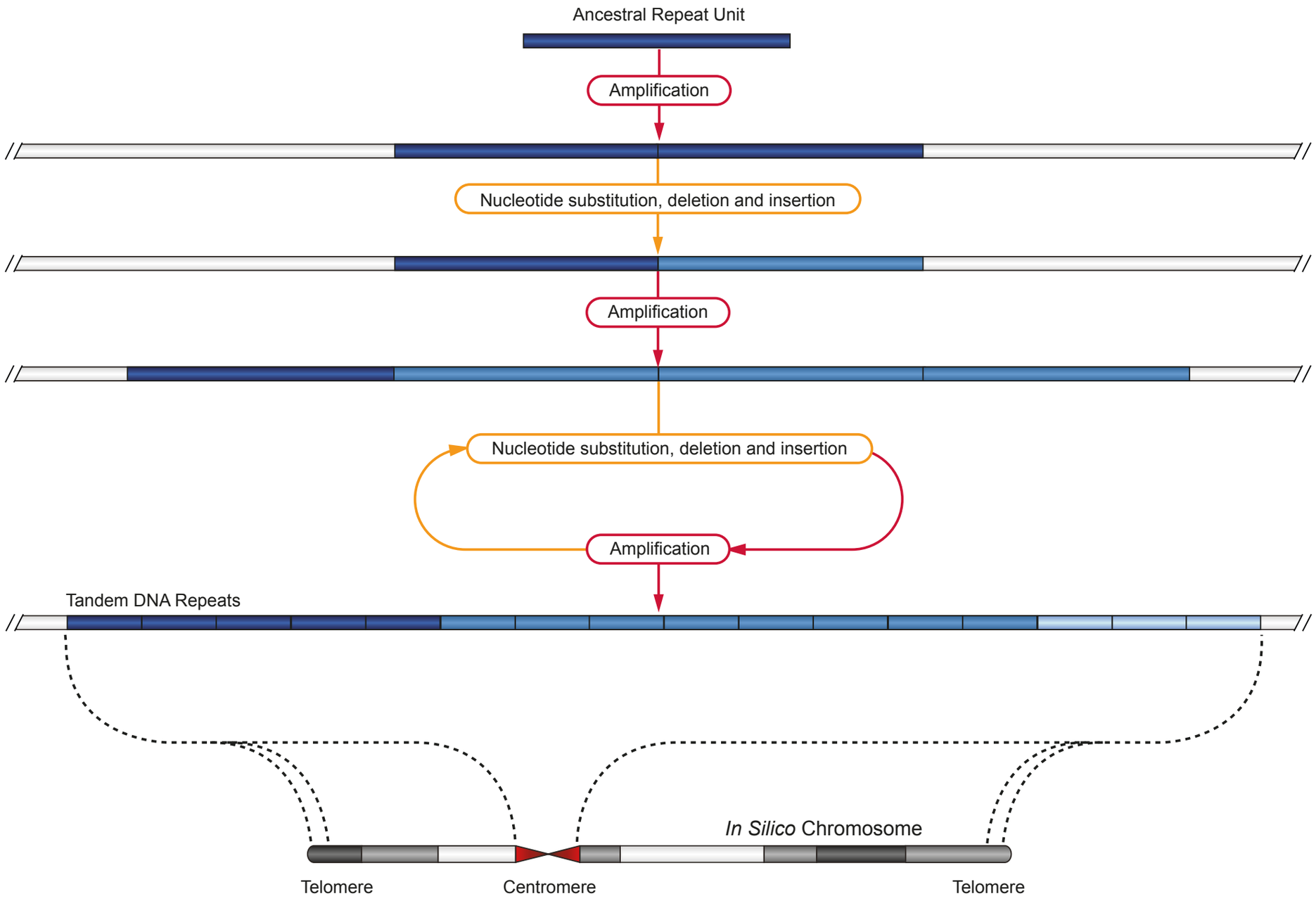
Structural Variation.



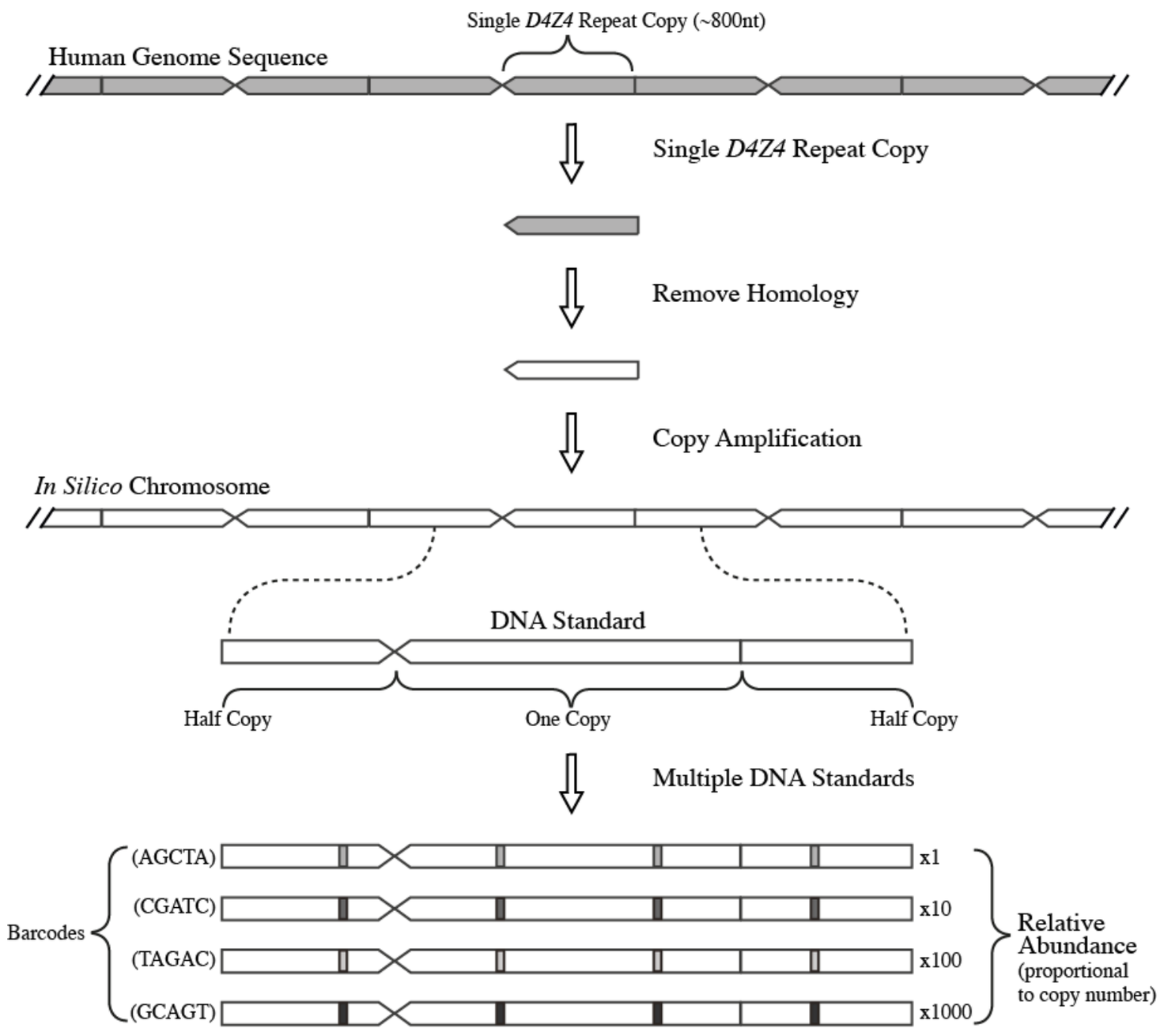
In silico chromosome / DNA standards emulate large-scale structural variation

Repeat DNA.

DNA standards establish a quantitive reference to measure repeat DNA (such as telomere length).

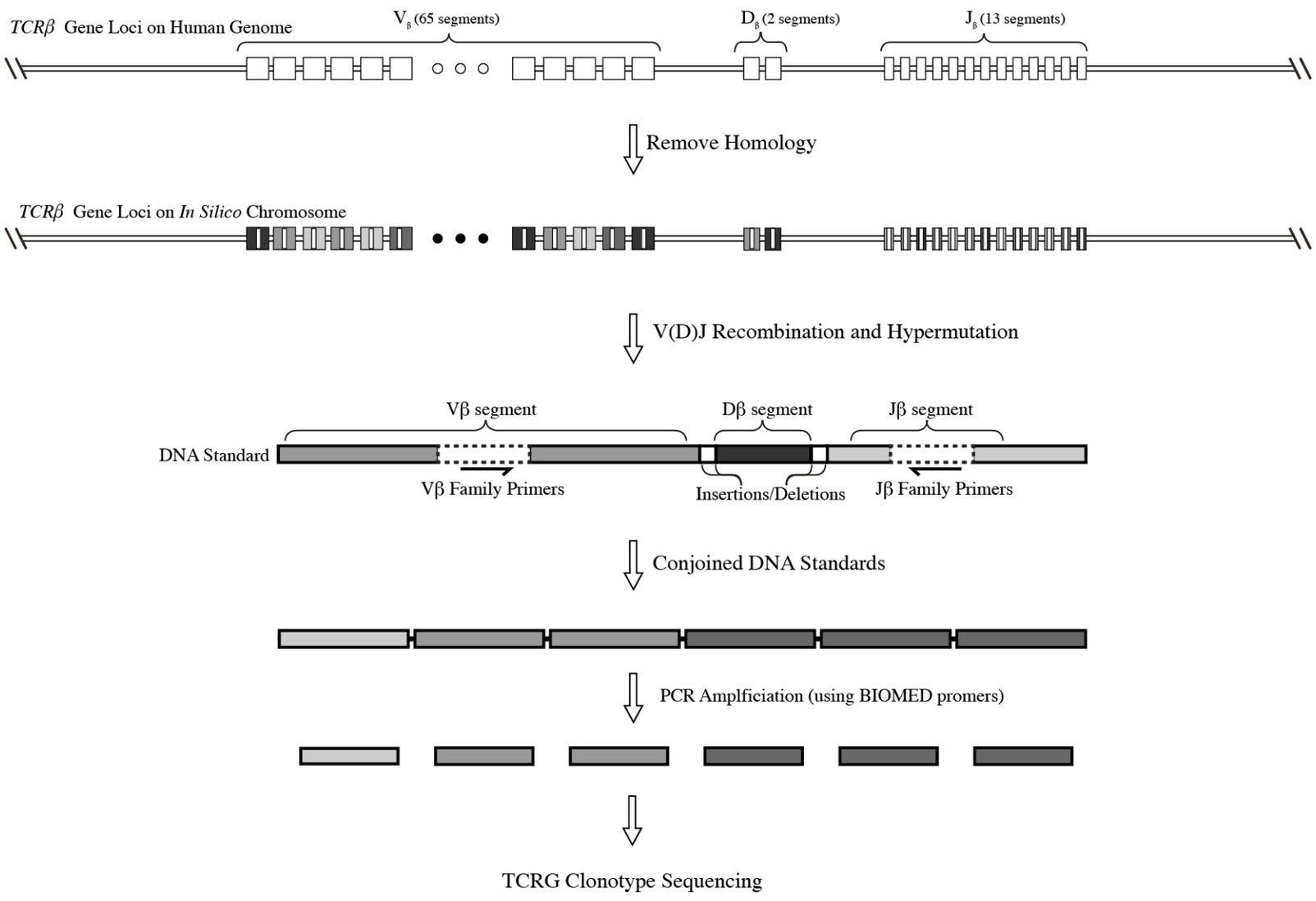


Clinical Structural Variation.



D4Z4 Repeat Contraction in Facioscapulohumeral Muscular Dystrophy

Immune Receptors.



1. Artificial immune receptors that retain complementary sequence to multiplex primers are encoded within in silk chromsome.
2. V(D)J recombination and somatic hypermutation are modeled with resultant clonotypes represented by DNA standards
3. Assess primer efficiency / PCR amplification and quantification during amplicon sequencing of immune receptor clonotypes

GARVAN INSTITUTE

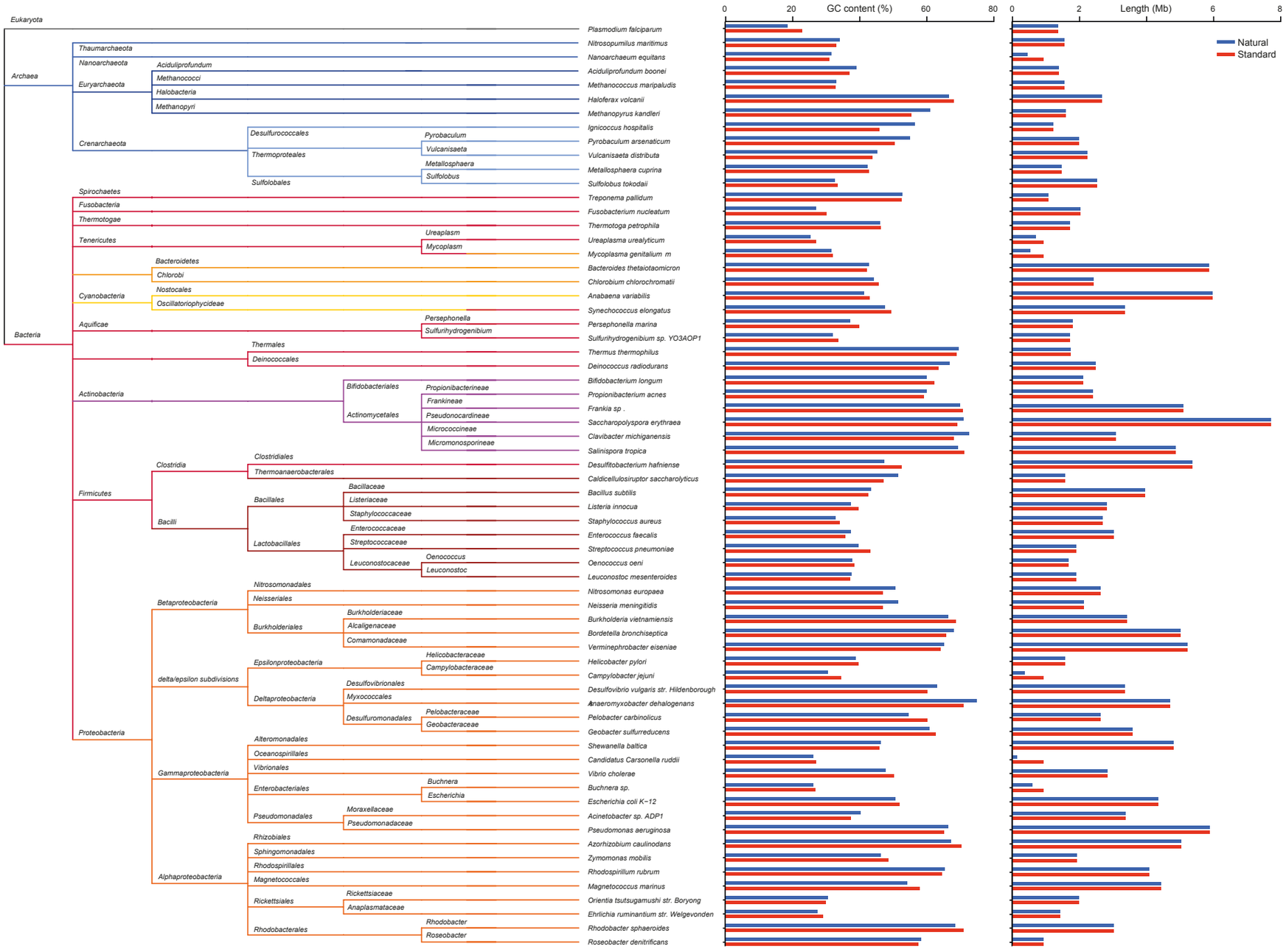
CONFIDENTIAL

Metagenomes.

Design of an *in silico* community of microbe genomes

In silico community represents diversity of microbe genomes (taxa, GC% and length)

In silico community represents diversity of microbe genomes (taxa, GC% and length)



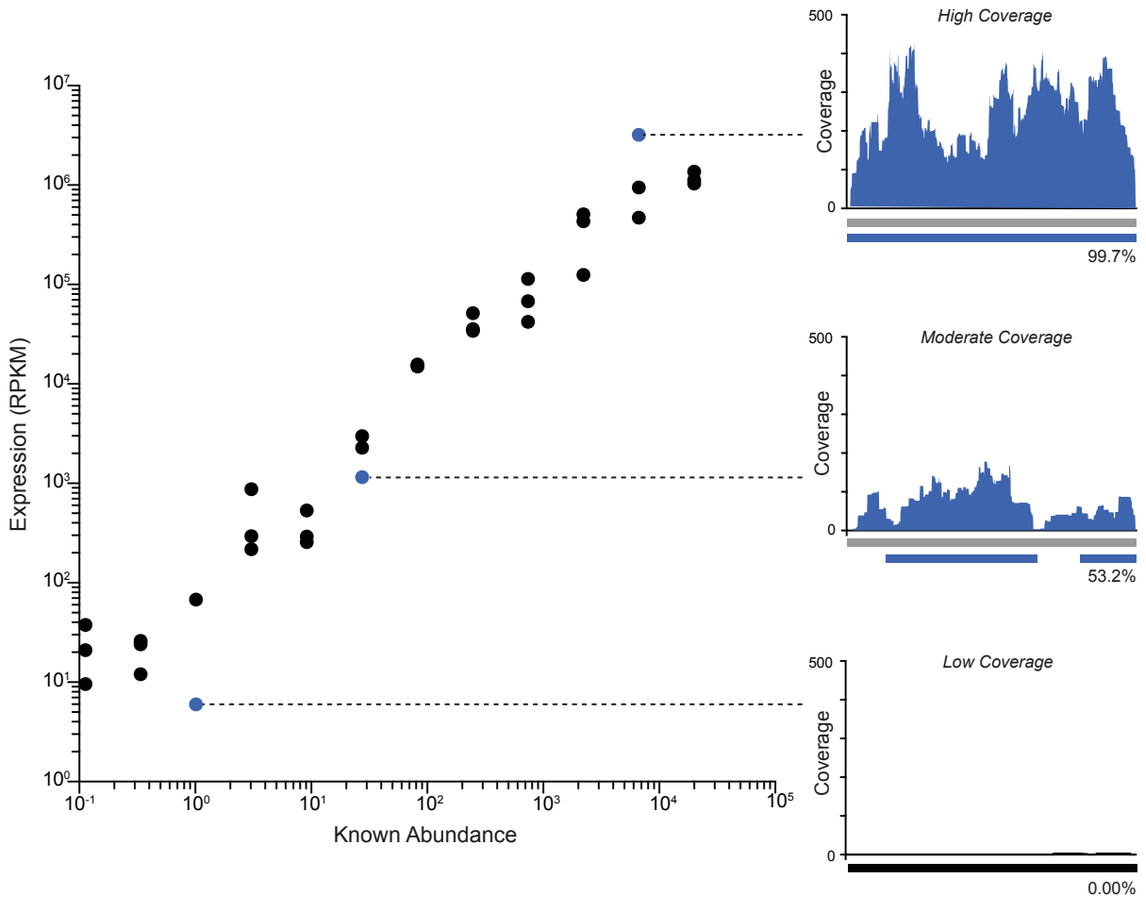
In silico community represented by DNA standards across 10^6 fold range in abundance.

GARVAN INSTITUTE

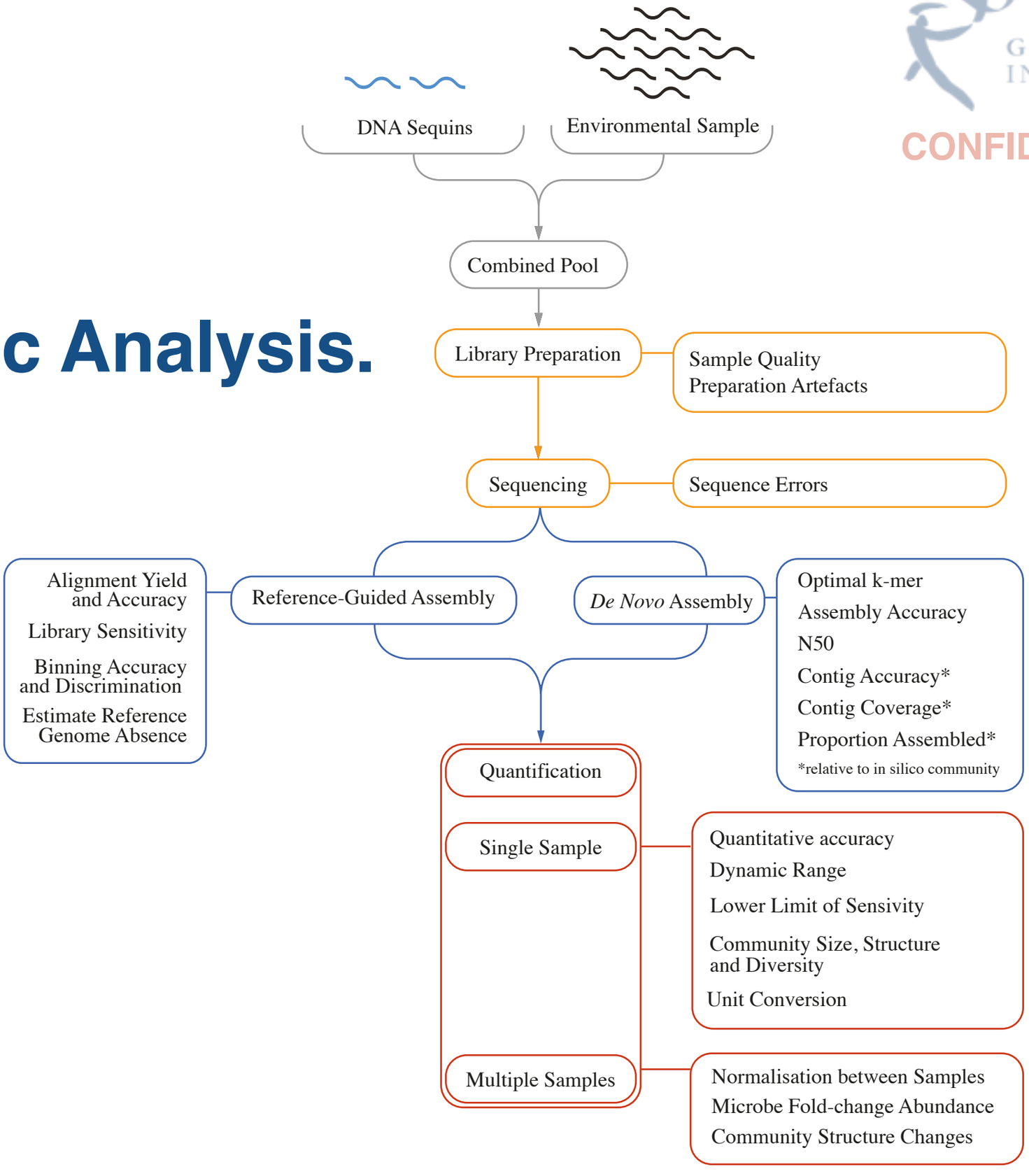
CONFIDENTIAL

Assess *De Novo* Assembly.

Impact of sequencing coverage and determine limit of sensitivity on assembly of microbe genomes.



Bioinformatic Analysis.



Assessment of metagenome sequencing

DNA Sample	Fecal 1	Fecal 2	Mangrove Site 1	Mangrove Site 2	Mangrove Site 3	Mangrove Site 4	Mangrove Site 5	Mangrove Site 6
DNA Standards Mixture	MixA	MixA	MixA	MixA	MixA	Mix B	Mix B	Mix B
Reads to Genome	225,662,643	229,258,092	76,069,281	64,245,984	80,676,540	70,760,900	70,054,850	74,436,488
Reads to *In Silico* Detection Limit	2,014,973	1,979,618	4,317,629	5,151,742	3,712,994	5,165,622	6,617,292	3,486,096
Fraction Dilution	0.0089	0.0086	0.0568	0.0802	0.0460	0.0730	0.0945	0.0468
***De Novo* assembly**								
Coverage	0.532	0.519	0.319	0.275	0.296	0.282	0.242	0.315
Nodes	23	78	32	26	604	434	23	38
N50	476	365	386	538	91	91	473	410
Maximum Contig Size	927	1007	1000	986	942	991	992	948
Total Bases in Assembly	0.029	0.022	0.030	0.033	0.024	0.024	0.023	0.046
Base								
Sensitivity	0.833	0.826	0.801	0.776	0.862	0.834	0.774	0.864
Specificity	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Quantification								
Correlation	0.970	0.954	0.961	0.960	0.968	0.951	0.961	0.954
Slope	1.041	1.037	1.061	1.093	1.026	1.164	1.177	1.125