Gene Doping: State of the art

Dereck de Mézquita[1]*, Shaïma Azouzi[2], Thomas Santiago[3]

Abstract
In 2004 the World Anti Doping Agency banned gene doping [1]; this along with the International Olympic Committee. Together they defined gene doping as "the transfer of genes, genetic elements, or the modulation of gene expression, which has the capacity to enhance athletic performance." It might still seem like science fiction, but every advancement in medical gene therapy brings us one step closer to *super humans*. We review the gene therapy technologies with the intent of doping. We cover the creation of "super athletes" with viral vectors such as adeno-associated virus, and detecting cheaters with PCR based techniques such as LAMP. We ponder future techniques for accurate detection.

Keywords
Gene doping, PCR based detection, Adeno-associated virus (AAV), Transgene.

[1] *Molecular Microbiology, University of Toulouse III, Toulouse, France*
[2] *Structural & Functional Biochemistry, University of Toulouse III, Toulouse, France*
[3] *Recombinant Protein Engineering & Expression, University of Toulouse III, Toulouse, France*
***Corresponding author**: dereck@demezquita.com

Contents

Introduction

Just as important as to the how is the why of doping. Doping has many goals but the primary will always be to improve performance; in this study sports. The key for improving performance in a sportsman has a basis in biology; as we are in fact biological machines. Modifying the production, regulation, and or action of certain proteins in an individual can have drastic effects on their biology and if done correctly even enhance their athletic performance.

In order to modify protein synthesis, or regulation in a lasting way one must target the genome. This is why scientists, transhumanists and even athletes have looked to gene therapy technologies. Several genes have been demonstrated to be essential in athletic performance, these are thus great theoretical targets for modification, see table 1.

Gene doping might still seem like a far-fetched and science fiction-like idea. Yet small genetic changes could yield orders of magnitude of improvement in performance, and in some cases are only a small modification away.

In the 1964 winter Olympic Games Finnish Skier Eero Mantyranta displayed an impressive performance and won two gold medals, it was later found that he had a small modification in his erythropoietin (EPO) receptor. In another case a child was found to have a mutation in a myostatin gene which caused muscular hypertrophy. These natural mutations offer us a blueprint to producing artificial versions in the lab. These prospects are exciting

Enhancement	Target gene(s)
Endurance	Erythropoietin (EPO)
	Vascular endothelial growth factor (VEGF)
	Hypoxia inducible factors (HIFs)
	Phosphoenolpyruvate carboxykinase (PEPCK)
Pain tolerance	Endorphins
	Downstream regulatory element antagonistic modulator (DREAM)
Speed & endurance	Angiotensin-converting enzyme (ACE)
	Peroxisome proliferator-activated receptor-δ(PPAR δ)
Speed	Actinin binding protein 3 (ACTN3)
Strength	Myostatin (inhibition)
	Follistatin
	Human growth hormone (hGH)
	Insulin-like growth factor-1 (IGF-1)

Table 1. Putative list of genes which could be modified for improvement in athletic performance.

for the creation of genetic therapies, and some also envision using such technologies in order to improve their current biology.

There are however, grave risks associated to gene modification, thus detection and prevention of cheaters remains a top priority for all regulatory bodies.

In 2004 the International Olympic Committee (IOC) along with the World Antidoping Agency (WADA) banned the practice of gene doping. They defined gene doping as "the transfer of genes, genetic elements, or the modulation of gene expression, which has the capacity to enhance athletic performance."

1. Transgene integration

The goal of genetic modification or gene doping is the transfer or modification of genetic sequences into/of the host. There are numerous methods in biology which allow for this transfer to occur. Indeed, various gene therapies are in clinical trials as of today; some diseases being worked on are: Parkinson's, Alzheimer's, arthritis, and even muscular dystrophy [2].

Various methods for the transfer of genetic material exist: viral vectors, lipid mediated vectors, hydrodynamic delivery, and gene guns [2]. *Ex vivo* approaches are possible but require specialised laboratories and careful monitoring of the subject [3]. However, it has been found that viral vectors still remain the most efficient and desired methods for transfer, they are in essence nature's gene gun [2]. Here will focus on viral vectors; notably adeno-associated virus.

1.1 Viral vectors and adeno-associated virus

Various viral vectors can be used in order to create mutants. Lentiviral vectors are based on the human immunodeficiency virus type 1 (HIV-1). Lentiviruses naturally have a narrow host cell spectrum [4]. In order to improve tropism those used for gene doping are pseudotyped with surface proteins from other viruses [4]. These viruses have been developed in such a way so that they are harmless to humans. To this end, self-inactive viruses (SINs) have been created by Ou *et al*. This reduces the probability that the integration of the transgene + viral genome will activate neighbour endogenous genes [4]. The creation of a lentivirus adapted for gene doping is similar to that of the method used for adeno-virus: undesirable viral genes are deleted and replaced by a suitable promoter and the transgene. The transfection of the athlete's cells can be done in vivo or in cell culture [4].

Daya *et al*, have found that adeno-associated virus (AAV) vectors remain among the most efficient and safest methods for genetic transfer/therapies [2]. There are a variety of advantages associated to an AAV vector. They present little to no toxicity [2], and can be engineered to integrate into the genome in a site specific manner if so desired [2]. Moreover AAVs distinguish themselves from other viral vectors by allowing for sustained and longterm gene expression after infection of cells [2]. Being that AAV does not integrate into the host genome as other retroviruses do it does not present a risk of random integration events [3]. AAV gene insertions can exist in a stable manner inside non-replicating cells for long periods of time, the amount of time is dependent on cell types and other factors [3].

AAV does require certain conditions in order to function properly as a gene therapy method. AAV is most efficient in postmitotic cells, because as previously men-

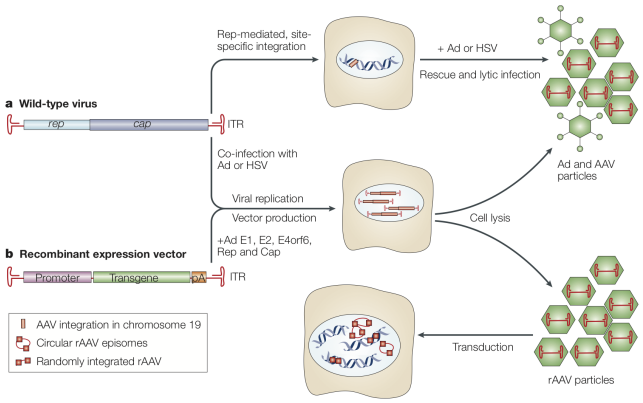

Figure 1. An illustration of rAAV production for use in transduction or transgene integration or expression.

tioned it does not integrate into the host genome [2]. Cell division would thus drive the loss or dilution of the episome. In contrast retroviruses require cell division in order to integrate and be stable [2].

The use of AAV as vectors may be relatively straight forward from a safety perspective but still presents some disadvantages. Even though AAV is mostly non-toxic and almost non immunogenic it does present some possibility of provoking an inflammatory response involving cytotoxic T cells [2]. The use of AAVs as a vector remains a complex reaction, dosage is very important. If a dose is too feeble then transgene expression could be negligible as was the case with Daya *et al.* and their dose of $2 \cdot 10^{12}\, vector\, genomes/kg$ [2]. However, too high a dose and one could possibly induce an immune response in certain cells [2]. Daya *et al.* found that there was a cytotoxic-T-lymphocyte (CTL) response to the AAV proteins in the liver [2]. In contrast Salganik used AAV at this same dose without too great an immune response [3].

The use of adeno-associated virus as a transgene vector requires reengineering the virus. In a recombinant adeno-associated (rAAV) instead of the virus having its natural genome, it contains the desired sequence and a promotor. There are a couple of steps necessary in this design process. First we must produce our viral package containing the desired DNA sequence. In order to do this we need an expression vector, *i.e.* cells that serve the sole purpose of producing our final product. The first step is to infect a cells with adeno virus. Adeno virus actually integrates itself into the genome of the organism. It contains all of the necessary sequences for integration but not for lysis or reproduction. A co-infection with wild-type adeno-associated virus allows for lytic infection and thus production of adeno virus, and adeno-associated virus virions (virial packets *ie* capsid + genomic sequence). A recombinant expression vector or a modified DNA sequence of adeno-associated virus is produced. This modified sequence replaces the viral DNA sequences with a chosen promotor and transgene; these are flanked by the same viral inverted terminal repeats (ITR).

A co-infection of these wild-type systems, and the recombinant expression vector yeilds, a mix of packaged recombinant adeno-associated viral particles (rAAV particles) and wild-type viral particles, see figure 1 for an illustration. Once separated these rAAVs can be used for the transduction of transgenes into a host.

2. Detecting gene doping

Gene doping detection in sports still remains theoretical as these rules have yet to be enforced. The principle is nonetheless important, thus foresight into possible detection methods is an important exercise.

Detection of gene doping in a host can be a tricky endeavour. If the goal of the genetic modification is to produce a protein inside the host then distinguishing the natural proteins from the transgenic proteins is tricky [1]. Detection of gene doping has been theorised to best work when targeting the genes themselves inside of a system. In order to transduce or introduce a set of new genes, into a system one has to simplify the modification as much as possible. This with two goals in mind, increasing efficiency as much as possible, and to reduce risk of complications. The longer the genetic sequences transduced the less the likely-hood of stable expression or even integration by the host [2].

Simplified forms of genes would include exon - exon fused genes for use in transgenic insertion. Note that in nature genes are often an alternating sequence of introns and exons. Introns are removed during splicing (after transcription) *i.e.* RNA maturation to form a single sequence of exon - exon RNA strands. Thus retro transcriptive polymerase chain reaction (rtPCR) assays are not possible as they would not be able to differentiate between the natural and exogenous materials. Logically the only place left to target if the subject indeed did use exon - exon joined transgenes is to test the DNA directly for these junctions.

Moreover, other tests can be done on the proteins directly, but these have been known to be tricky and

are still in development. It is still unknown whether we could have a definitive method for differentiating exogenous proteins from their natural counterparts.

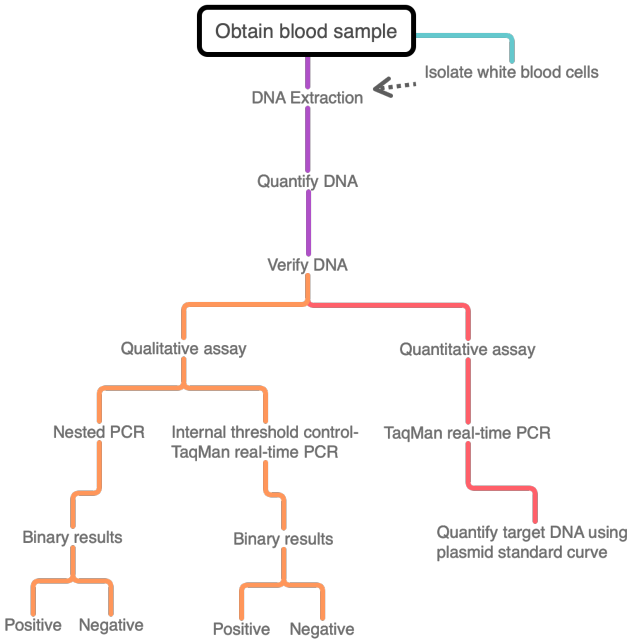

Figure 2. Flow chart for detection, there are two main branches for detection qualitative assays or quantitative.

Detection of transgenes in a person requires multiple steps, depending on the required data one choses either a qualitative assay or quantitive. Since anti-doping agencies are mostly interested in the absence of any modification a binary qualitative assay such as loop mediated isothermal amplification (LAMP) should be enough. A general flow-chart for sampling and detecting transgenes in samples, see figure 2.

2.1 PCR based detection

2.1.1 PCR

Classic polymerase chain reaction (PCR) allows for the qualitative detection of transgenes; this is based on exon - exon junctions. Exon - exon junctions are not natural in nature and thus can be a target for detection by use of overlapping primers.

It is theoretically possible to detect gene doping with conventional PCR. The results obtained by PCR can be visualised by electrophoretic gel after a bath in ethidium bromide, see figure 3. If the transgene is present, it will be amplified and a large quantity of DNA molecules with equivalent size will be produced, and can be visualised as a single band. Only one primer pair would be used corresponding to the junctions as well; the results seen in figure 3 show detection of bacterial contaminants in the food industry. Wang *et al.* are working for developing LAMP and other PCR based techniques for detection and prevention of food contamination [5].

This assay would yield binary results, positive for a transgene or negative (depending on how the primers were designed). Note that the possibility for false positives is still there. Depending on transgene design, primer design, and the host's own genome the transgenes could be undetectable.

2.1.2 LAMP

Loop mediated isothermal amplification (LAMP) is a PCR-based method. Unlike a conventional PCR, 4 primers are used which hyridise 6 total regions [6]. As with PCR detection primers correspond to exon - exon junctions on the gene of interest. In addition, only two of the primers are integrated in the amplified molecules. The other have a role only for the first amplification cycles. LAMP also uses BST polymerase for its amplification, this polymerase has two particularities: it can amplify (elongate) in an isothermal state (about 60 °C), and posses a strand displacement property; which allows this isothermal reaction [7].

The second major difference with a conventional PCR is that LAMP is a nested PCR. In other words, the product of a cycle is a matrix for the next cycle. It is therefore not necessary to start from a large amount of sample DNA. In addition, the product obtained in LAMP is a single strand, containing many copies of the target, and not separate DNA fragments containing single copies of the gene. This feature is one of the advantages of LAMP. Indeed, the results obtained with LAMP are on/off. In other words, if the transgene is present in the athlete's body, and only in this case, an amplification will happen. The amplification is very sensitive and specific, it can only take place in the presence of the transgene, moreover the analysis of the results is very simple.

The amplification product can be analysed by an electrophoretic gel, on which a smear appears, see figure 3 [5]. This smear corresponds to all products of the LAMP amplification, intermediate or final giant molecules. Note that the LAMP reaction happens asynchronously. Detection can also be done with LAMP in a simple tube by adding calcein molecules which are chelated with Mn^{2+} initially. As the reaction progresses the amplification produces pyrophosphate which

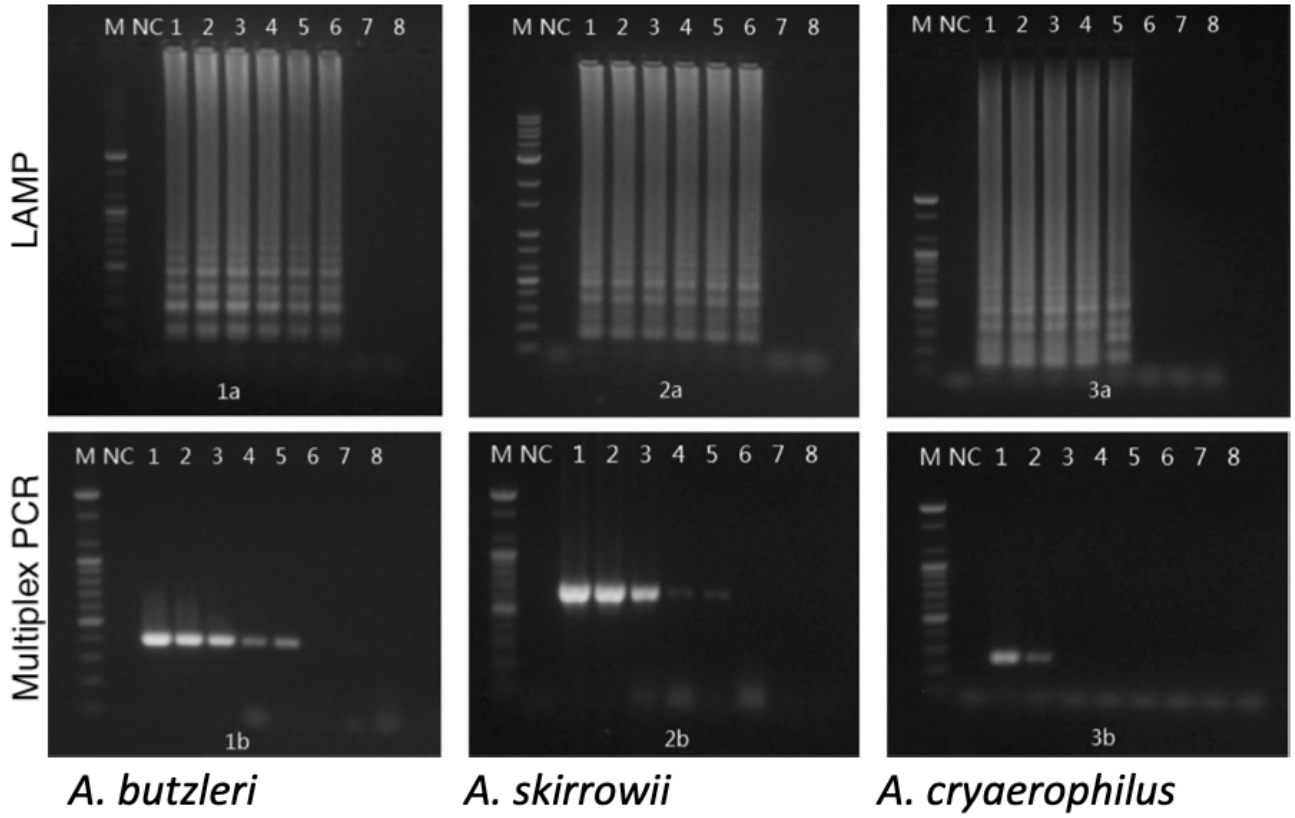

Figure 3. Electrophoretic gels of both LAMP and multiplex PCR products. Lanes as follows: M 100 bp DNA ladder, NC negative control, L1 $2X10^5$ CFU, L2 $2X10^4$ CFU, L3 $2X10^3$ CFU, L4 $2X10^2$ CFU, L5 $2X10^1$ CFU, L6 2 CFU, L7 $2X10^{-1}$ CFU, L8 $2X10^{-2}$ CFU. Detection in 99/120 assays: 20 at 0h, 24 at 3h, 27 at 6h, 28 at 24h after contamination with the LAMP method. Detection in 51/120 assays: 0 at 0h, 6 at 3h, 18 at 6h, 27 at 24h after contamination with the multiplex PCR method.

chelates both Mg and Mn forming $Mg_2P_2O_7$ and $Mn_2P_2O_7$ which precipitate [6]. Calcein bound to Mn^{2+} is quenched and does not fluoresce yet its form bound to Mg^{2+} does. The equilibrium is modified as amplification occurs and the Mn^{2+} form of calcein is favoured this allows there to be fluorescence and thus indicates a positive result *i.e.* the gene of interest was detected [6]. These results can even be seen under normal sunlight and all can be done in a single tube reaction thus vastly reducing the amount of necessary materials and cost, see figure 4 [6].

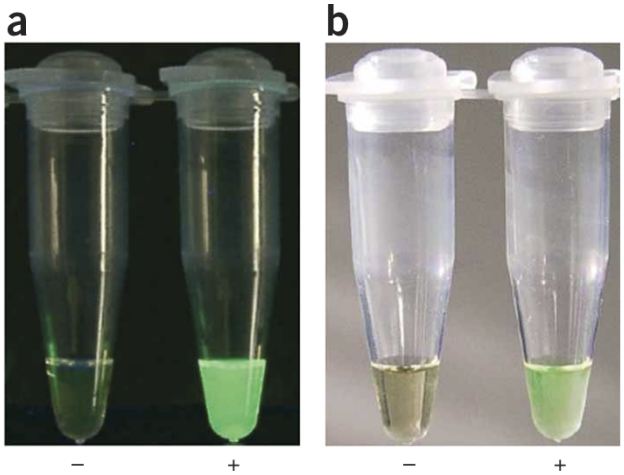

Figure 4. Detection with LAMP PCR. Left under 365 nm handheld UV light, right under normal sunlight. Detection +, negative result -. Positive detection can be determined visually even under normal sunlight.

2.2 Other methods

Although PCR-derived methods are the most promising, other techniques have been considered. Those are based on different characteristics inherent to gene doping. For example, some have tried to detect an increased immune response following infection with the adeno-associated virus [1]. This detection would be made via the analysis of antibodies produced by the host. However, this

method is unreliable because AAV is a widespread virus, and a large proportion of the population may have already come into contact with it [1]. The possibility of obtaining a false positive is therefore non-negligible.

Structural analysis of the proteins produced by the endogenous gene versus the exogeneous is another approach worth trying. Although the structure of the two proteins is mostly identical, it is possible that differences reside in post-transcriptional modifications [1]. However, again, this method is unreliable since these post-transcriptional modifications may vary within the same individual. Indeed, taking the example of EPO, it is known that it can be found in different glycosylated forms in an organism. It would therefore be difficult to ensure that the glycosylated form observed is not in fact, a natural form in this individual [1]. Other methods have been proposed but as above, they are highly constrained; often requiring large databases, to which reliable comparisons can be made.

3. Prospects

3.1 Hurdles and problems

Every detection method seems to have a downfall, see figure 5. Depending at which level one attacks is the results one gets. This is mostly due to a lack of data and technology. Expression profiling requires standards/reference databases. Immune response assays can yield false positives. Structural studies have some way to go as we do not fully understand the interplay between protein synthesis and trans-protein production; post translational modifications of such are rather unknown. Moreover, tissue specific promotors can be easily changed and require random testing/databases. Finally DNA barcodes would require mass coordination between antidoping agencies and large amounts of sequencing. Salamin *et al.* propose an interesting method of detection using LAMP but then again it depends on how one designs the primers for hybridisation *i.e.* the target.

It is for all these reasons that it can be said that the detection of gene doping is a major problem, still far from being solved. The proposed techniques often pose an increased risk of false positives, or even non-detection. Techniques based on PCR analyses appear to be the most promising as they target the genetic material at the origin of the proteins and are inexpesive. Since, these methods are based on the detection of exon-exon junctions, they can thus be circumvented.

For these reasons we assert that the most reliable detection would be an association of several methods. The first would be a genomic analysis of the expression profile, in order to detect a potential overexpression of a gene. This genomic analysis should be supplemented by a metabolic analysis. Indeed, the over-expression of a protein is certainly not inconsequential, and it is possible that some metabolic pathways may be disturbed. A fine knowledge of the metabolic pathways as well as the relative amount of metabolites present in the organism at an instant, in a given situation. This would imply a a huge amount of data. Such work would require several decades, but could be one of the only options for testing whether gene doping has taken place.

3.2 Machine learning and big data

It seems that the ideal approach would be a wholistic one, *i.e.* to compare an athlete's expression profile to a database and model; here we would need and leverage data issue from various sources. It is impossible to compare the expression profile of two different individuals without analysis and a model. There are too many variables: epigenetic expression differs from one individual to another. The solution would be comparison of the results with a theoretical model.

A database must be built. It should be representative of a large proportion of the population in general. We should sample the population generally since athletes can come from any background. We should account for different types of profiles, including variables such as: sex, age, fitness, lifestyle *etc*. This type of database does not exist.

Theoretically these expression profiles, could be built through microarray analysis, and other metabolic assays. Testing would rely on comparing the athlete's expression profile (amount of mRNA produced) with a standard. The reason we would be interested in their mRNA expression profile is because we want to see the whole expression profile, we theorise that inserting a transgene or other modification into the mix could change said expression profile on all levels: transcription, translation, degredation *etc*.

Something similar was studied by Deo *et al*: machine learning for application in medicine [8]. One important bottle-neck in producing exploitable data is the robustness of experimental controls, this is important to the integrity of the data internally but also for the ensemble of data produced. Although, data fidelity might be an

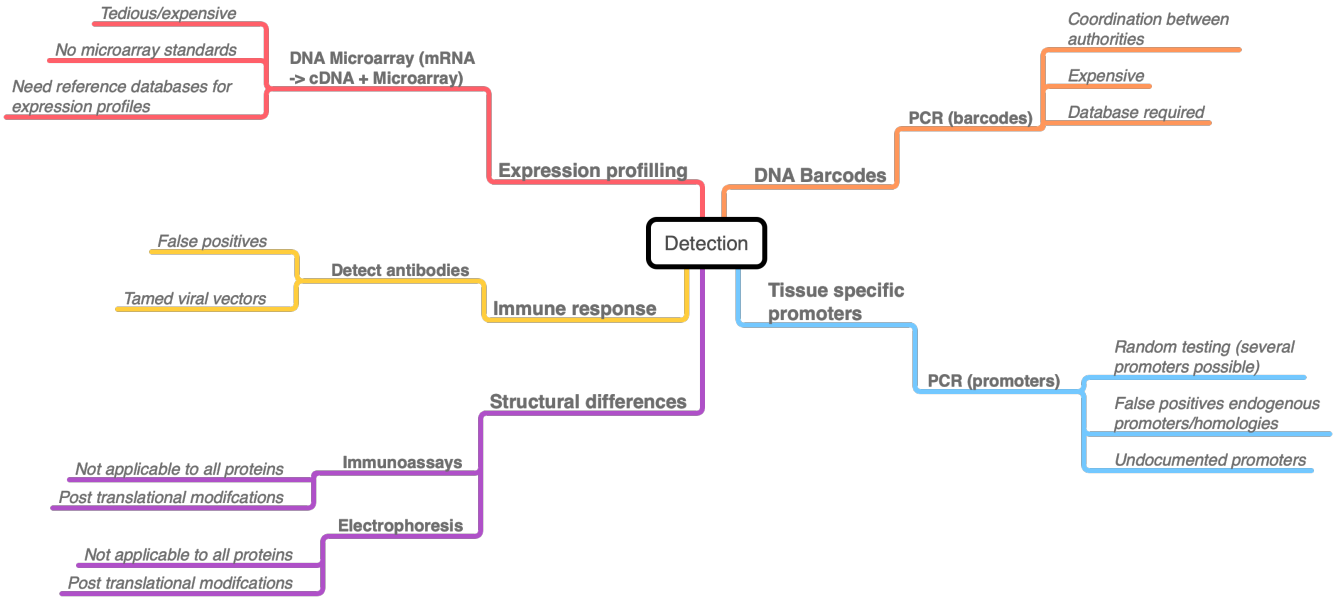

Figure 5. A map of different possible detection strategies and specific methods; each method has advantages and disadvantages to it.

issue a possible solution is drowning out un-exploitable data by volume; the machine learning algorithm if implemented correctly should be able to discriminate and exploit the data [8].

This database and model can be used to determine presence of doping or not. The other complimentary assays could thus be used to score the result. This data could be pooled together with that of other candidates and ran through statistical analysis and machine learning algorithms to determine a "doping profile". Were machine learning to be used it seems that k-nearest neighbour algorithms or decision trees would be most apt to create this model as per Deo *et al.* in their application of machine learning to medicine [8].

This profile would be built over time and once a significant model is obtained it could be used to analyse metabolic data. Having a large data base and a developed algorithm we should be able to determine if an athlete has a transgene or not [8].

3.3 Ethics

Although theoretically a lot of methods exist for catching gene dopers, they remain intrusive; as they require physical sampling. These physical samples in theory do not pose too much of an issue in modern sports as most athletes undergo some sort of doping tests. Yet here we propose collecting, storing, and leveraging genomic data. This would require large data bases of whole genomes of real persons. Security would be a great issue, and requiring such testing and forfeiture of data might be seen as unethical. Accurate detection is a must when there is a risk of damaging an athlete's career and reputation.

3.4 Hurdles & circumvention

As gene editing technologies advance more and more, the prospects of producing gene doped athletes that present wild-type like genomes become more likely. If athletes wishing to dope find a way to integrate an exact copy of the endogenous gene, then it will be simply impossible to differentiate it from the native gene by a PCR method. At this point it would be nearly impossible to detect transgenes in cheaters.

It is theorised that full wild-type genes could be inserted into a host (exons + introns) by use of adeno-associated virus [2]. Daya *et al.* proposed that a large gene sequence including introns + exons could be introduced as a transgene by use of adeno-associated virus or another virus vector by splitting the transgene into a multitude of parts to be reassembled by recombination inside the host.

Acknowledgments

We would like to thank Pr. Dumas and Pr. Poupot for their guidance and assistance during this course. It was a unique learning experience and a fun one at that. We greatly enjoyed our time.

References

[1] Olivier Salamin, Tiia Kuuranne, Martial Saugy, and Nicolas Leuenberger. Loop-mediated isothermal amplification (LAMP) as an alternative to PCR: A rapid on-site detection of gene doping. *Drug Testing and Analysis*, 9(11-12):1731–1737, 2017.

[2] Shyam Daya and Kenneth I Berns. Gene Therapy Using Adeno-Associated Virus Vectors. 21(4):583–593, 2008.

[3] Max Salganik, Matthew L Hirsch, and Richard Jude Samulski. HHS Public Access. 3(4), 2015.

[4] Li Ou, Michael J. Przybilla, Brenda L. Koniar, and Chester B. Whitley. Elements of lentiviral vector design toward gene therapy for treating mucopolysaccharidosis I. *Molecular Genetics and Metabolism Reports*, 8:87–93, 2016.

[5] Xiaoyu Wang, Dong Joo Seo, Min Hwa Lee, and Changsun Choi. Comparison of Conventional PCR, Multiplex PCR, and Loop Mediated Isothermal Amplification Assays for Rapid Detection of Arcobacter Species. 52(2):557–563, 2014.

[6] Norihiro Tomita, Yasuyoshi Mori, Hidetoshi Kanda, and Tsugunori Notomi. Loop-mediated isothermal amplification (LAMP) of gene sequences and simple visual detection of products. *Nature Protocols*, 3(5):877–882, 2008.

[7] Nobuyuki Amino Tsugunori Notomi, Hiroto Okayama, Harumi Masubuchi, Toshihiro Yonekawa, Keiko Watanabe and Tetsu Hase. Loop-mediated isothermal amplification of DNA Tsugunori. *Oxford University Press*, 12(28), 2000.

[8] Rahul C Deo. Machine Learning in Medicine HHS Public Access. *Circulation*, 132(20):1920–1930, 2015.